Schätzung des Stichprobenfehlers mit Stata: eine Einführung mit Beispielen zum Campus File Mikrozensus 2002

Schimpl-Neimanns, Bernhard

Veröffentlichungsversion / Published Version
Monographie / monograph

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
GESIS - Leibniz-Institut für Sozialwissenschaften

gesis
Leibniz-Institut
für Sozialwissenschaften

METHODENBERICHTE 2009|02

Schätzung des Stichprobenfehlers im Mikrozensus mit Stata - Eine Einführung mit Beispielen zum Campus File Mikrozensus 2002

Bernhard Schimpl-Neimanns

GESIS-Methodenberichte

GESIS - Leibniz-Institut für Sozialwissenschaften
German Microdata Lab (GML)
Postfach 12 21 55
68072 Mannheim
Telefon: (0621) 1246 - 263
Telefax: (0621) 1246 - 100
E-Mail: gml@gesis.org

ISSN: 1865-7567 (Print)
ISSN: 1865-7575 (Online)
Herausgeber, Druck
und Vertrieb: GESIS - Leibniz-Institut für Sozialwissenschaften
Lennéstraße 30, 53113 Bonn

Zusammenfassung

Für Schätzungen von statistischen Kennwerten der Grundgesamtheit aus der Stichprobe muss das Stichprobendesign berücksichtigt werden. Für diese Zwecke enthalten die Mikrozensus Scientific Use Files entsprechende anonymisierte Informationen zur Schichtung und Klumpung. Diese Informationen sind ebenfalls Bestandteil des für den Einsatz in der Lehre entwickelten Campus Files des Mikrozensus, das als Public Use File allgemein zugänglich ist. Im Wesentlichen können die für das Campus File entwickelten Verfahren auch auf die Scientific Use Files des Mikrozensus übertragen werden. Der Bericht zeigt die Möglichkeiten, die es für die Schätzung des Stichprobenfehlers mit den anonymisierten Files und mit dem Statistikprogramm Stata gibt. Beispielhaft werden die Schätzer und ihre Standardfehler für Gesamt-, Anteils- und Mittelwerte sowie Differenzen als interessierende Parameter vorgestellt. Hierbei werden sowohl Schätzungen bei freier Hochrechnung bzw. Designgewichtung als auch bei gebundener Hochrechnung, d. h. mit Anpassung der Mikrozensusergebnisse an demografische Populationsverteilungen durchgeführt. Ergänzend wird beschrieben, wie bei statistischen Modellen vorgegangen werden kann, um evtl. durch das Stichprobendesign bedingte Modellverletzungen zu beheben. Die Programme sind im Anhang dokumentiert.

Abstract

To estimate statistical population parameters from survey data the design features of the sampling strategy must be incorporated into the analysis. For this purpose the Scientific Use Files (SUF) of the German Mikrozensus contain anonymised design information, such as variables identifying strata and clusters. This information is also part of the so-called Mikrozensus Campus File (CF), which was developed for academic teaching and statistical training, and which is accessible as a public use file. By and large, analysis procedures for the Mikrozensus Campus File can be applied to Scientific Use Files as well. This report demonstrates the possibilities of estimating sampling errors using the Mikrozensus 2002 CF and the statistical computing package Stata. Techniques covered include variance estimation of totals, ratios, means and differences in parameters when using design weights as well as post-stratification. Additionally, it is illustrated how to estimate statistical models if the usual assumptions are possibly violated due to the sampling design. The Stata programs are documented in the appendix of this report.

Inhaltsverzeichnis

1 Einleitung

Im Mikrozensus werden mit einem Stichprobenumfang von einem Prozent der Personen und Haushalte Informationen über die demografische, soziale und wirtschaftliche Struktur der Bevölkerung erhoben. Infolge des hohen Auswahlsatzes erlauben diese Daten differenzierte Analysen bei geringem Stichprobenfehler. Bei der Berechnung von Stichprobenfehlern ist allerdings zu berücksichtigen, dass der Mikrozensus keine einfache Zufallsstichprobe, sondern eine geschichtete Klumpenstichprobe ist. Das pragmatische Ziel dieses Berichtes ist es, zu zeigen, wie der Stichprobenfehler im Mikrozensus mit Stata geschätzt werden kann. Eine frühere Fassung lag als Arbeitspapier zum GESIS-Workshop „Stichprobendesign und Hochrechnungsverfahren im Mikrozensus – Praktische Übungen zum Thema Hochrechnung und Gewichtung“ vor, der in Zusammenarbeit mit dem Statistischen Bundesamt am 12. und 13. Juni 2008 in Mannheim durchgeführt wurde. Da die darin dargestellten Analysen aber für einen größeren Nutzerkreis von Interesse sein können, wurde der Bericht überarbeitet und geringfügig ergänzt.[1]

Während noch vor einigen Jahren in den üblicherweise in der Forschung benutzten Statistikprogrammen kaum Verfahren zur Schätzung des Stichprobenfehlers bei geschichteten Klumpenstichproben zur Verfügung standen, bieten mittlerweile alle gängigen Statistikpakete solche Prozeduren an.

Abbildung 1: Leistungsumfang verschiedener Statistikprogramme bei der Schätzung von Stichprobenfehlern

Sampling design feature	Stata	SUDAAN	WesVar	SAS	SPSS
probability weight	√	√	√	√	√
stratification	√	√	√	√	√
PSUs	√	√	√	√	√
levels of sampling	multi	two	one	one	two
post-stratification	√	√	√		
replicate weights	√	√	√		
subpop for all procs	√	√	√		√

Quelle: Wells 2007.

SPSS wird in den Sozialwissenschaften am häufigsten genutzt und wäre somit für die praktischen Übungen mit den Mikrozensusdaten zu präferieren. Jedoch sind die „SPSS Complex

[1] Für hilfreiche Anmerkungen zu früheren Fassungen danke ich Wolf Bihler, Siegfried Gabler, Ulrich Kohler, Ulrich Pötter und Julia Schroedter.

Samples Procedures“ wenig bekannt und nicht im Standardpaket enthalten, sondern müssen extra erworben werden. Deshalb wird das Programm Stata verwendet, das als eines der ersten Statistikpakete vielfältige Standardprozeduren für die Analyse von Stichproben mit komplexen Designs angeboten hat.

Die Forschung kann den Mikrozensus in Form einer faktisch anonymisierten 70-Prozent-Substichprobe (Scientific Use File) nach Abschluss eines Bereitstellungsvertrages nutzen. Für die Umsetzung und Erprobung der Varianzschätzungen wird das öffentlich zugängliche Campus File des Mikrozensus 2002 eingesetzt.[2] Es enthält als eine 3,5 %-Wohnungssubstichprobe des Mikrozensus elementare Informationen zum Stichprobendesign, die für eine exemplarische Varianzschätzung benötigt werden.

Tabelle 1: Vergleich ausgewählter Randverteilungen im Campus File (CF) und im Scientific Use File (SUF) des Mikrozensus 2002

Bevölkerungsgruppe (Filter)	Wohnberechtigte Bevölkerung		Bevölkerung in Privathaushalten		Bev. in PrivatHH u. Hauptwohnsitz	
Daten	CF	SUF	CF	SUF	CF	SUF
Hochrechnung / Gewichtung[1]	d	d	v751g	v751	v750g	v750
Auswahlbezirke (n ungewichtet)	10.707	45.058				
Wohnungen (hochger. in 1.000)[2]	32.440	32.434	35.133	35.120		
(n ungewichtet)	11.354	227.037				
Haushalte (hochger. in 1.000)	33.300	33.305	38.751	38.719		
(ungewichtet)	11.655	233.135	11.655	233.135		
Personen (hochger. in 1.000)	71.820	71.868	82.745	82.756	81.612	81.615
(ungewichtet)	25.137	503.075	24.881	498.075	24.534	491.073
			in Prozent			
Geschlecht (EF32)						
1 männlich	48,1	48,2	48,7	48,7	49,0	49,0
2 weiblich	51,9	51,8	51,3	51,3	51,0	51,0
Staatsangehörigkeit (EF52)						
1 Deutscher	94,0	93,7	91,8	91,8	91,4	91,4
2 Ausländer aus EU-Staaten	1,4	1,7	1,9	2,1	2,0	2,3
3 Ausländer aus Nicht-EU-Staaten	4,6	4,6	6,3	6,1	6,6	6,3
Erwerbstyp (EF504)						
1 Erwerbstätige	44,6	44,4	44,9	44,7	44,8	44,6
2 Erwerbslose (ILO-Def.)	4,0	4,1	4,2	4,2	4,2	4,3
3 Sonstige Erwerbslose	0,7	0,7	0,7	0,7	0,7	0,7
4 Nichterwerbspersonen	50,7	50,8	50,2	50,4	50,3	50,4

1) Gewichtungsvariablen im CF: Designgewicht: d = 1/(0,01*0,035); Gebundene Hochrechnung Wohnungsfaktor: ef761g; Haushalts-/Familienfaktor: ef751g; Person: ef750g; Gewichtungsvariablen im SUF: Designgewicht: d = 1/(0,01*0,70); Gebundene Hochrechnung: Wohnungsfaktor: ef761; Haushalts-/Familienfaktor: ef751; Person: ef750.

2) Wohnungen in Wohngebäuden. Die gebundene Hochrechnung basiert auf dem Wohnungsfaktor.

Quelle: Campus File und Scientific Use File Mikrozensus 2002; eigene Berechnungen.

[2] Für weitere Informationen siehe die WWW-Seiten der Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder: www.forschungsdatenzentrum.de/bestand/mikrozensus/cf/2002/index.asp.

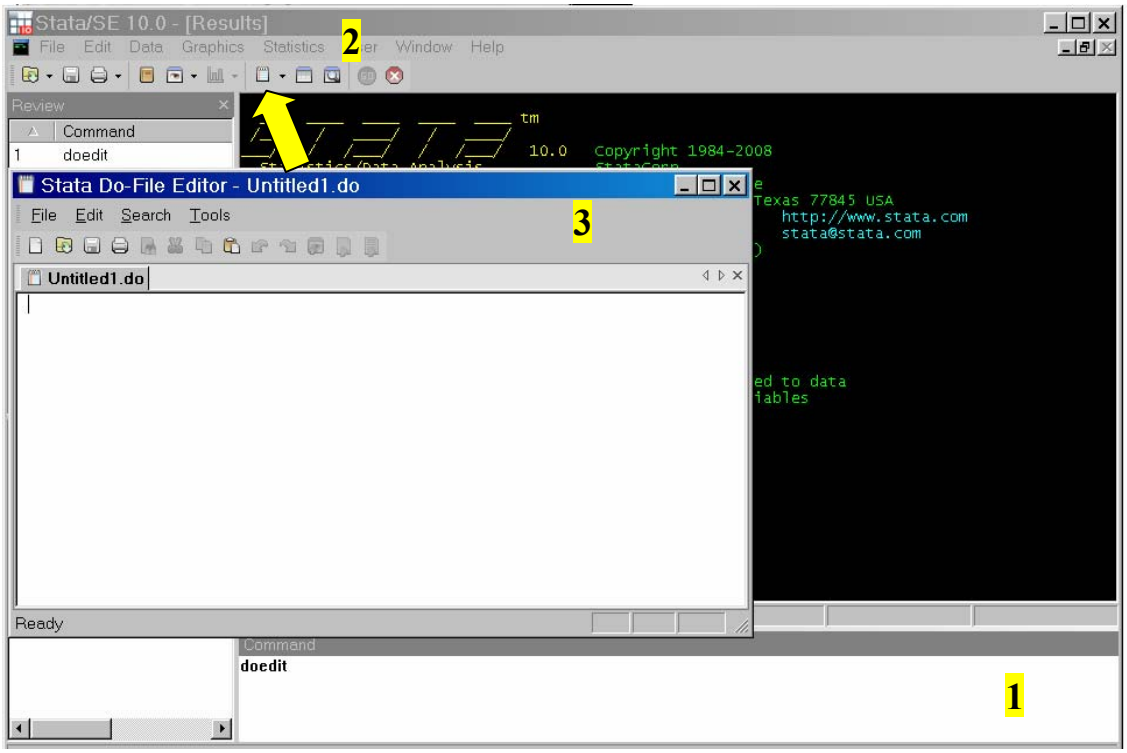
Im Wesentlichen können die am Beispiel des Campus File entwickelten Stata-Programme auch auf die Scientific Use Files übertragen werden, wenn man entsprechende Korrekturen für den höheren Auswahlsatz von 70 Prozent vornimmt. Dennoch ist zu beachten, dass die charakteristischen Designelemente des Original-Mikrozensus aufgrund des geringen Stichprobenumfangs bzw. des Ziehungsverfahrens des Campus Files nur näherungsweise abgebildet werden können. Während im Scientific Use File ein Auswahlbezirk (Klumpen) durchschnittlich 5,2 Haushalte (= 233.135 / 45.058) enthält, umfasst er im Campus File nur etwa 1,1 Haushalte (= 11.655 / 10.707; siehe Tab. 1). Der mit dem Campus File ermittelbare Klumpeneffekt spiegelt deshalb im Wesentlichen die Klumpung auf Haushalts- bzw. Wohnungsebene wider. Die Merkmalsverteilungen sind weitestgehend vergleichbar, sodass zumindest für ausreichend besetzte Merkmale plausible und mit Veröffentlichungen vergleichbare Ergebnisse ermittelt werden können.

Im Folgenden werden die wichtigsten Stata Befehle für die Schätzung von Gesamtwerten (Totals), Verhältnis- (Ratios) und Mittelwerten (Means) bei freier Hochrechnung bzw. Designgewichtung und gebundener Hochrechnung, d. h. mit Anpassung der Mikrozensusergebnisse an demografische Populationsverteilungen beispielhaft erläutert. Abschließend wird beschrieben, wie bei statistischen Modellen vorgegangen werden kann, um evtl. durch das Stichprobendesign bedingte Modellverletzungen zu beheben. Die jeweils in Stata verwendeten Formeln werden nur kurz dargestellt; zu Herleitungen wird auf Lehrbücher verwiesen (z. B. Cochran 1972, Krug et al. 2001; Särndal et al. 1997). Um die Anwendung der Programme auch für Personen, die bisher vorwiegend mit SPSS gearbeitet haben, zu erleichtern, wird zu Beginn skizziert, worin sich SPSS und Stata hinsichtlich der Syntax unterscheiden.

2 Stata Grundlagen

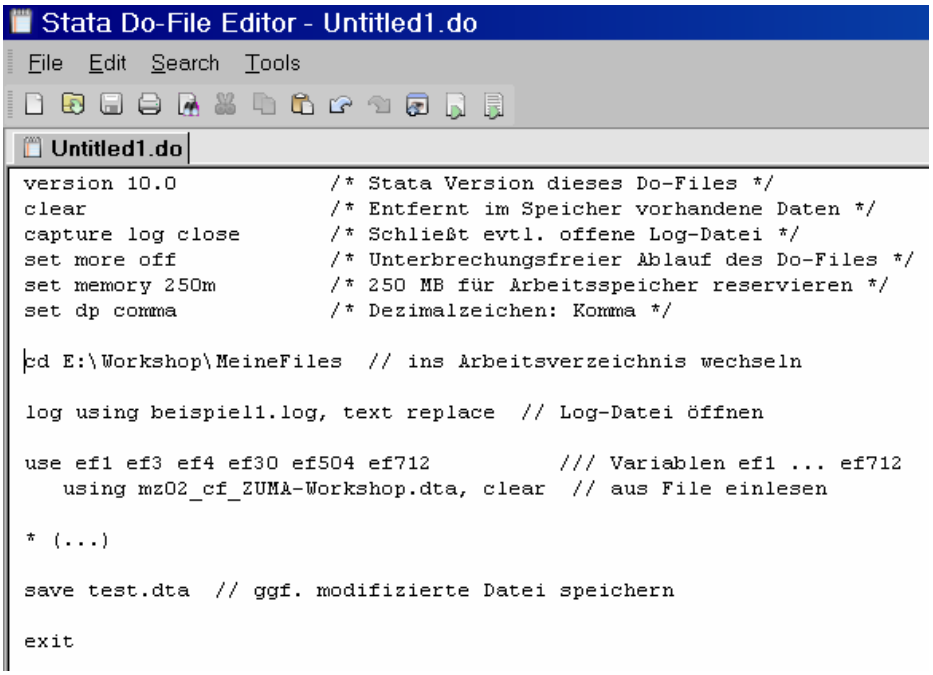
Stata-Kommandos können (1) interaktiv im Fenster „Command“, (2) menügesteuert oder (3) per Syntax-Datei (Do-File) zur Ausführung gebracht werden. Da die Speicherung der Kommandos in einer Syntax-Datei für die Replikation von Analysen vorteilhaft ist, konzentriert sich dieser Überblick darauf. Die Syntax kann mit dem Do-File-Editor von Stata, aber auch einem beliebigen Texteditor geschrieben werden.

Der Do-File-Editor kann über das Menü (siehe unten: Pfeil) oder mit dem Befehl `doedit` gestartet werden. Mit `doedit Dateiname` wird eine bereits existierende Datei geöffnet.



Ein typisches Do-File (Dateiname.do) kann wie folgt beginnen und entweder menügesteuert mit „Tools“ oder nach dem Speichern der Datei („File“) und Wechsel in das „Command“-Fenster mit dem Befehl `do Dateiname.do` gestartet werden. Bevor das Stata-File geladen wird,[3] wird eine Log-Datei zur Protokollierung benannt.

[3] In SPSS würde man äquivalent zum Stata-Kommando `use ef1 (...) using Dateiname.dta` das Datenfile mit `GET FILE='Dateiname.sav' /KEEP = ef1 (...)` einlesen.



Kommentare können wie oben gezeigt mit einem * am Zeilenanfang oder zwischen den Zeichen „ /* “ und „ */ “ eingeschlossen werden. Mit „ /// “ wird erreicht, dass der Befehl in der nächsten Zeile fortgesetzt wird. Die nach „ // “ bzw. „ /// “ stehenden Zeichen werden nicht als Befehl interpretiert, sodass dieser Platz zur Kommentierung verwendet werden kann.

Zu beachten ist, dass in Stata Groß- und Kleinschreibung unterschieden werden. Wenn also z. B. der Variablenname ef1 ist, muss das list-Kommando `list ef1` lauten, da die Großschreibung von EF1 zu einem Fehler führt.

Erläuterungen zu den Kommandos und evtl. Optionen können mittels `help Kommando` abgerufen werden. Einen ersten Überblick zu den Daten geben die folgenden Kommandos:

Stata	SPSS	Kurzbeschreibung
`notes`	`display document`	Anmerkungen zu Daten und Variablen zeigen
`ds`	`display names`	Auflisten aller Variablen
`list`	`list variables`	Ausgabe von Variablen und Fällen
`browse`		Auflisten von Variablen und Fällen im Tabellenformat („Datentabelle“)

Beispielsweise können mit dem folgenden Befehl die ersten zehn Personen bzw. fünf Haushalte nach den Identifikatoren Bundesland (ef1), Auswahlbezirksnummer (ef3), Haushaltsnummer (ef4), Personennummer (ef5) und den Variablen Zahl der Personen im Haushalt (ef500), Stellung innerhalb des Haushalts (ef507), Alter (ef30) und Geschlecht (ef32) gelistet werden:

```
list ef1 ef3 ef4 ef5 ef500 ef507 ef30 ef32 in 1/10, noobs ///
     sepby(ef1 ef3 ef4) string(10)

-------------------------------------------------------------------
ef1        ef3  ef4  ef5  ef500     ef507         ef30          ef32
-------------------------------------------------------------------
Schleswig-H  1    1    1  4 Pers..  [1] Bezugs..  37 Jahre  Männlich
Schleswig-H  1    1    2  4 Pers..  [2] Ehegatte  36 Jahre  Weiblich
Schleswig-H  1    1    3  4 Pers..  [3] (Schwi..   5 Jahre  Weiblich
Schleswig-H  1    1    4  4 Pers..  [3] (Schwi..   3 Jahre  Weiblich
-------------------------------------------------------------------
Schleswig-H  2    1    1  2 Pers..  [1] Bezugs..  52 Jahre  Männlich
Schleswig-H  2    1    2  2 Pers..  [2] Ehegatte  51 Jahre  Weiblich
-------------------------------------------------------------------
Schleswig-H  3    1    1  1 Person  [1] Bezugs..  33 Jahre  Weiblich
-------------------------------------------------------------------
Schleswig-H  4    1    1  2 Pers..  [1] Bezugs..  37 Jahre  Männlich
Schleswig-H  4    1    2  2 Pers..  [2] Ehegatte  40 Jahre  Weiblich
-------------------------------------------------------------------
Schleswig-H  5    1    1  1 Person  [1] Bezugs..  39 Jahre  Männlich
-------------------------------------------------------------------
```

2.1 Fall- und Variablenselektionen

Bei der Fall- und Variablenselektion sowie bei der Variablenkonstruktion sind folgende logische und relationale Operatoren relevant:

und	`&`	oder	`|`
nicht	`!` `~`		
größer als	`>`	kleiner als	`<`
größer gleich	`>=`	kleiner gleich	`<=`
gleich	`==`	nicht gleich	`!=` `~=`

Zu beachten ist, dass bei Berechnungen und Sortierungen die Zeichen für fehlende Werte (`.`, `.a`, ... , `.z`) programmintern auf $+\infty$ gesetzt werden, so dass z. B. die Bedingung `if (z >= 2)` auch für fehlende Werte wahr ist. Es ist deshalb zu empfehlen, den zulässigen Wertebereich immer explizit anzugeben.

Zum permanenten Löschen von Variablen aus dem Arbeitsfile können die Kommandos `keep Variablenliste` und `drop Variablenliste` verwendet werden (SPSS: `delete variables Variablenliste`).

Analog dazu können permanente Fallselektionen mittels `keep if` und `drop if` erreicht werden, z. B. stehen nach dem Befehl `keep if ef506==1` nur noch Personen zur Verfügung, die zur Bevölkerung in Privathaushalten zählen. (SPSS: `select if (ef506=1)`).

Temporäre Fallselektionen können in Stata mittels `if`–Anweisung umgesetzt werden, z. B. bei der Tabellierung:

Stata	SPSS

`tabulate ef525 if ef506==1`	`temporary.`
	`select if (ef506=1).`
	`frequencies variables = ef525.`

2.2 Variablen modifizieren und neu erstellen

Neue Variablen werden mit `generate` (Abkürzung: `gen`) erzeugt. Bereits existierende Variablen können mit `replace` verändert werden. Dies kann auch mittels Rekodierung einer gegebenen Variable in eine neue Variable erreicht werden.

Stata	SPSS

`gen v72 = ef72`
`replace v72 = 5 if (ef72==6)`
`recode ef72 (6 = 5), gen(v72) ///`
`copyrest`
	`compute v72 = ef72.`
	`if (ef72 = 6) v72 = 5.`
	`recode ef72 (6=5) (else=copy)`
	`into v72.`

Stata bietet mit den Kommandopräfixen `by` und `bysort` insbesondere für hierarchisch strukturierte Daten wie die des Mikrozensus Möglichkeiten der zeilenübergreifenden Konstruktion von Variablen auf der Ebene von Wohnungen, Haushalten, Familien und Lebensgemeinschaften, bei denen die Ordnungsnummern dieser Einheiten genutzt werden.[4] Mit dem Befehl `egen` stehen erweiterte Funktionen von `generate` bereit, um z. B. die Haushaltsgröße bzw. die Zahl der Personen in Privathaushalten (ef521) selbst zu ermitteln.[5]

Stata	SPSS

`bysort ef1 ef3 ef4: ///`
`gen v521 = _N if ef506==1`
	`recode ef506 (1=1) (3=0) into p.`
	`sort cases by ef1 ef3 ef4.`
	`aggregate outfile = *`
	`/mode = addvariables`
	`/presorted`
	`/break ef1 ef3 ef4`
	`/v521 = sum(p).`

`egen v521 = total(ef506==1), ///` `by(ef1 ef3 ef4)`	

Mit dem ersten Stata-Befehl (`bysort (...) if ef506==1`) wird Personen in Gemeinschaftsunterkünften (ef506=3) aufgrund der if-Bedingung in der Ergebnisvariablen v521 der Missing-Code „`.`“ zugewiesen, mit der zweiten Alternative `egen` wird Null vergeben. Für fehlende Werte können mit Stata generelle („`.`“) oder erweiterte Missing-Werte („`.a`“, ..., „`.z`“) zugewiesen werden. Das von den Forschungsdatenzentren bereitgestellte Stata File ent-

[4] Siehe auch weitere Beispiele für SPSS und Stata unter http://www.gesis.org/Dauerbeobachtung/GML/Service/Mikrodaten-Tools/Bandsatz96_04/index.htm.

[5] `_N` ermittelt die Summe der Einheiten.

hält für eine Reihe von Variablen fehlende Werte („.“), die für eine korrekte Varianzschätzung in gültige Werte rekodiert werden müssen. Dies kann effizient mit `mvencode` geschehen und mit `mvdecode` rückgängig gemacht werden.[6]

Stata	**SPSS**

`recode v521 (. = 0)`
`recode v521 (. = .a "Gem.Unt.")`
`mvencode _all, mv(-1)`
`mvdecode _all, mv(-1)`
	`recode v521 (sysmis = 0).`
	`missing value v521 (0).`

2.3 Variablen und Value Labels

Standardgemäß werden Variablen Label beispielsweise zugewiesen mit

```
label var v521 "Zahl der Personen in Privathaushalten"
```

Außer der Zuweisung von Value Labels im `recode` Kommando (s. o.) werden für die Vergabe zwei Schritte benötigt. Mit dem Befehl `numlabel` können den Value Labels die numerischen Werte vorangestellt werden.

```
label def v521 0 "Gemeinschafts-/Anstaltsunterkunft" ///
               1 "1 Person" {...} 9 "9 Personen"
label val v521 v521
numlabel v521, add mask("[#] ") force detail
```

2.4 Deskriptive Auswertungen

Die meisten deskriptiven Auswertungen und statistischen Modelle können in Stata auch mit gewichteten Daten durchgeführt werden. Stata unterscheidet vier Gewichtungstypen (s. u.). Mittels `help weight` kann man erfahren, welche Gewichte bei den einzelnen Kommandos zulässig sind und welche Optionen (z. B. Prozentuierung) es gibt.

Gewicht	**Erläuterung** (Ausgabe von `help weight`**)**
`fweights`	**Frequency** `fweights` indicate replicated data. The weight tells the command how many observations each observation really represents. `fweights` allow data to be stored more parsimoniously. The weighting variable contains positive integers. The result of the command is the same as if you duplicated each observation however many times and then ran the command unweighted.
`pweights`	**Sampling** `pweights` indicate the inverse of the probability that this observation was sampled. Commands that allow `pweights` typically provide a cluster() option. These can be combined to produce estimates for unstratified cluster-sampled data. If you must also deal with issues of stratification, see [SVY] survey.

[6] In SPSS können zusätzlich zu allgemein fehlenden Werten (sysmis) auch numerische Werte als fehlend (missing) behandelt werden. Dies ist in Stata nicht möglich.

Gewicht	**Erläuterung** (Ausgabe von `help weight`)
`aweights`	**Analytic** `aweights` are typically appropriate when you are dealing with data containing averages. For instance, you have average income and average characteristics on a group of people. The weighting variable contains the number of persons over which the average was calculated (or a number proportional to that amount).
`iweights`	**Importance weights**. This weight has no formal statistical definition and is a catch-all category. The weight somehow reflects the importance of the observation and any command that supports such weights will define exactly how such weights are treated.

Hier sind beispielhaft einige Kommandos für einfache Auswertungen:

Stata	**SPSS**
Einfache Häufigkeitsauszählungen	
`tab1 ef32 ef35`	`frequencies /variables ef32 ef35.`
Bivariate Kreuztabellen	
`tab ef32 ef35`	`crosstabs /tables ef32 by ef35.`

`tab ef32 ef35 [iw = v750g], /// row col`	`weight by v750g.`
	`crosstabs /tables ef32 by ef35`
	`  /cells count row column.`

Mehrdimensionale Kreuztabellen	

`table ef504 ef35 ef32`
 	
`bysort ef32: tab ef504 ef35`
	`crosstabs /tables ef504 by ef35`
	`  by ef32.`

Mittelwerte (z. B. Haushaltsgröße)	

`mean ef521 if ef506==1`	`temporary.`
	`select if (ef506=1).`
	`means tables = ef521`
	`  /cells mean semean.`

Gesamtwerte (z. B. Zahl der Haushalte)	

`gen hh = ef507==1`
`total hh if ef506==1`
	`recode ef507 (1=1) (else=0)`
	`  into hh.`
	`temporary.`
	`select if (ef506=1).`
	`descriptives variables = hh`
	`  /statistics = sum.`

Verhältniswerte (z. B. Frauenerwerbsquote)	

`gen ep  = ef504<=2 & ef505<=2`
`gen bev = ef505<=2`
`ratio ep/bev if ef32==2`
	`if ef504<=2 & ef505<=2 ep=1.`
	`if ef505<=2 bev=1.`
	`recode ep bev (missing=0).`
	`temporary.`
	`select if ef32=2.`
	`ratio statistics ep with bev`
	`  /print=mean CIN(95).`

3 Der Stichprobenplan des Mikrozensus und das Survey-Kommando im Überblick

Die obigen Beispiele beziehen sich ausschließlich auf die (implizite) Standardannahme, dass die Daten aus einer einfachen Zufallsstichprobe stammen. Dies trifft jedoch auf die anonymisierten Daten nicht zu, die als Substichproben des Mikrozensus wie die Originaldaten als geschichtete Klumpenstichproben gekennzeichnet sind. Wird das Stichprobendesign nicht berücksichtigt, d. h., geht man von der Annahme einer einfachen Zufallsstichprobe aus, werden i. d. R. die Standardfehler unterschätzt, die Konfidenzintervalle sind zu klein und Hypothesentests sind fälschlicherweise eher „statistisch signifikant“.

3.1 Stichprobendesign des Mikrozensus

Die folgende Zusammenstellung gibt einen Überblick zum Stichprobendesign des Mikrozensus ab 1990.

Tabelle 2: Überblick zum Erhebungsdesign des Mikrozensus 1990-2004, der Scientific Use Files 1996-2004 und des Campus File 2002

Stichprobeneigenschaften	**Mikrozensus ab 1990 – Originalmaterial**
Erhebungseinheiten	Haushalte, Personen
Auswahlgrundlage	Alte Bundesländer: Volkszählung 1987; Neue Bundesländer/Ost-Berlin (ab 1991): Bevölkerungsregister Statistik 1990 Aktualisierung der Stichprobe unter Berücksichtigung der Neubautätigkeit
Auswahlverfahren	Einstufig geschichtete Klumpenstichprobe
• Schichtung	Bundesland, Regierungsbezirk, Anpassungsschicht, Regionalschicht, Gebäudeschicht
• Auswahleinheiten	Primäreinheiten (PSUs): Auswahlbezirke PSUs sind Klumpen von i.d.R. zusammenliegenden Gebäuden bzw. Gebäudeteilen. Ein Auswahlbezirk verbleibt vier Jahre in der Stichprobe. In jedem Jahr scheidet ¼ der Auswahlbezirke aus (rotierendes Panel). Bildung der PSUs der Grundauswahl nach der Gebäudegröße (Gebäudeschicht): 1-4, 5-10, 11+ Wohnungen, Gemeinschaftsunterkünfte. Ein Auswahlbezirk der Grundauswahl (ohne Gemeinschaftsunterkünfte) umfasst durchschnittlich neun Wohnungen. Modifikationen der Gebäudeschicht bei Neubauauswahl: 1-4, 5-8, 9+ Wohnungen; Richtgröße jeweils sechs Wohnungen pro PSU.
• Auswahltechnik	Grundauswahl: 1. Sortierung der PSUs nach regionaler Schichtuntergruppe, Kreis, Gemeindegrößenklasse und PSU-Nr. 2. Zusammenfassung von jeweils 100 aufeinander folgenden PSUs zu einer Zone. 3. Zufällige Zuordnung der PSUs einer Zone zu den Zahlen 0-99 (=“Stichprobennummer“). Anschließend Zusammenfassung der PSUs mit gleicher Stichprobennummer in 100 1%-Stichproben. 4. Zufällige Zuordnung von je 4 aufeinander folgenden Zonen

Stichprobeneigenschaften	**Mikrozensus ab 1990 – Originalmaterial**
	(= "Block") zu den Zahlen 1-4 zur Zerlegung der 1%-Stichproben in 4 Rotationsviertel á 0,25%. 5. Ermittlung von 20 1%-Stichproben durch zufällige Ziehung der Ordnungsnummern der Stichproben. Zufällige Ziehung der ersten Stichprobe für 1990. Die Grundauswahl (1-5) kann zusammenfassend als uneingeschränkte Zufallsauswahl beschrieben werden. Aktualisierung/Neubauauswahl: 6. Sortierung nach Aktualisierungsjahr und regionaler Kennung. Systematische Auswahl mit Zufallsstart.
Stichprobenumfang	Ca. 390.000 Haushalte; ca. 830.000 Personen (2002; hochgerechnete, an die Bevölkerungsfortschreibung angepasste Fallzahlen; s. u.)
Auswahlsatz	1 Prozent
Hochrechnung	Zweistufiges Verfahren: 1. Kompensation der bekannten Ausfälle auf Haushaltsebene in 449 regionalen Untergruppen für jeweils 19 Merkmalskombinationen. 2. Anpassung der Stichprobenergebnisse an Eckzahlen aus der laufenden Bevölkerungsfortschreibung auf der Ebene von 132 regionalen Anpassungsschichten. Die Anpassungsklassen werden dabei gebildet durch die Angaben über die Zahl von Deutschen und Ausländern in der Gliederung nach Geschlecht. Die Anpassung für Soldaten und Wehrpflichtige erfolgt getrennt auf Regierungsbezirks- bzw. Landesebene auf Basis von Bestandsmeldungen des Verteidigungs- bzw. Innenministeriums. Die endgültigen Hochrechnungsfaktoren ergeben sich aus der Multiplikation des haushaltsbezogenen Kompensations- und des personenbezogenen Anpassungsfaktors.

Stichprobeneigenschaften	**Scientific Use Files des Mikrozensus 1996-2004 Unterschiede zum Originalmaterial**
Auswahlverfahren	Mehrstufiges Ziehungsverfahren (idealtypisch): 1. Stufe wie in MZ-Originalauswahl (s. o.) 2. Stufe: Ziehung der 70%-Substichprobe von Haushalten wie folgt [Abweichend davon wurde in den Scientific Use Files 1998 und 2002 eine 70%-Substichprobe von Wohnungen gezogen.]
• Schichtung durch Anordnung	a) Sortieren der Datensätze nach Bundesland, Regierungsbezirk, Gemeindegrößenklasse, Zahl der Personen in Privathaushalten, Auswahlbezirks-Nummer, Nummer des Haushalts im Auswahlbezirk. (Datensätze für leer stehende Wohnungen und ausgefallene Haushalte werden vor der Sortierung gelöscht.)
• Auswahltechnik	Systematische Zufallsauswahl: b) Nach der Sortierung werden Haushalte fortlaufend nummeriert. Hierbei werden Personen in Gemeinschaftsunterkünften wie Einpersonenhaushalte behandelt. (c) Ziehen/Löschen aller Sätze, deren letzte Platzziffer der Haushaltsnummer nicht einer von sieben ganzzahligen Zufallszahlen (z = 0, 1, 3, 4, 6, 7, 8) entspricht. D. h. Übernahme von 70 % der Haushalte in die Substichprobe. d) Die Auswahlbezirks- und Haushaltsnummern werden nach der Substichprobenziehung neu fortlaufend nummeriert (EF3, EF4).
Stichprobenumfang	227.037 Wohnungen in Wohngebäuden, 233.135 Haushalte; 503.075 Personen (2002; nicht hochgerechnete Fallzahlen)
Hochrechnung	Das Scientific Use File wurde nicht extra an die laufende Bevölkerungsfortschreibung angepasst. Es enthält die MZ-Hochrechnungsfaktoren für

Stichprobeneigenschaften	**Scientific Use Files des Mikrozensus 1996-2004** **Unterschiede zum Originalmaterial**
	Personen und Haushalte/Familien, die Ergebnisse des oben beschriebenen zweistufigen Verfahrens abbilden (EF750ff.).
Stichprobeneigenschaften	**Campus File Mikrozensus 2002** **Unterschiede zum Scientific Use File**
Auswahlverfahren**:** Schichtung und Auswahltechnik	Systematische Zufallsauswahl: Sortierung analog zur Ziehung des Scientific Use Files (s. o.), jedoch werden für die Substichprobenziehung (c) die letzten drei Platzziffern der Wohnungsnummer verwendet. Mit zufälligem Startwert werden bei vorliegender Sortierfolge 35 von jeweils 1.000 Wohnungen zufällig ausgewählt.
Stichprobenumfang	11.354 Wohnungen in Wohngebäuden, 11.655 Haushalte; 25.137 Personen (nicht hochgerechnete Fallzahlen)
Hochrechnung	Das Campus File enthält die Original-MZ-Hochrechnungsfaktoren für Personen und Haushalte/Familien und Wohnungen sowie speziell für das Campus File konstruierte Hochrechnungsfaktoren (EF750G, EF751G, EF761G), die eine direkte Hochrechnung auf ein Prozent der Population ermöglichen sollen.

3.2 Das Survey-Kommando

Mit Stata ist es möglich, eine Reihe von Stichprobendesigns zu definieren. So können z. B. mehrstufige Klumpendesigns bei deskriptiven Analysen und einer Vielzahl statistischer Modelle berücksichtigt werden. Die zentralen Designelemente werden in Stata mit `svyset` definiert und unten kurz erläutert:[7]

```
svyset [psu] [weight] [, design_options options]
  design_options:    strata, fpc
  options:           Poststratification (poststrata, postweight) {...}
```

- **Primäreinheit** (primary sampling unit; **PSU**): Die der Stichprobenziehung zugrunde liegenden Auswahlbezirke (Klumpen) umfassen als künstlich abgegrenzte Flächenstücke mehrere in der Regel benachbarte Gebäude oder Gebäudeteile. Alle Wohnungen, Haushalte und Personen eines gezogenen Auswahlbezirks werden als sekundäre Auswahleinheiten (SSUs) erfasst.
- **Sekundäreinheiten**: Im Fall mehrstufiger Stichproben können analog zur einstufigen Auswahl die Design-Optionen für zweite und ggf. folgende Stufen hinter den Zeichen „||“ definiert werden. Im anonymisierten Mikrozensus betrifft dies insbesondere die Ziehungswahrscheinlichkeit von Wohnungen bzw. Haushalten.
- **Gewichtung** (**weight**): Im einfachsten Fall wird zur Hochrechnung bzw. Gewichtung der Stichprobe das Designgewicht, d. h. der Kehrwert der Inklusionswahrscheinlichkeiten verwendet (pw). Es können auch andere Gewichte eingesetzt werden. Zur gebundenen Hochrechnung siehe „Poststratification“. Für die Ergänzungs- und Zusatzprogramme (v. a.

[7] Siehe dazu viele Anwendungsbeispiele im Web unter www.ats.ucla.edu/stat/stata/topics/Survey.htm, die neben Stata auch für SAS und weitere Statistikprogramme angeboten werden.

Merkmale der EU-Arbeitskräfteerhebung), die bis 2004 als Substichprobe durchgeführt wurden, liegen allerdings die auf Regierungsbezirksebene variablen Auswahlsätze von 0,4 %, 0,6 %, 0,8 % oder 1 % in den anonymisierten Daten aus Datenschutzgründen nicht vor, sondern es ist nur der im Bundesgebiet durchschnittliche Auswahlsatz von 0,45 % bekannt.

- **Schichtung** (**strata**): Alle Einheiten der Population sind eindeutig einer Schicht *h* zugeordnet. Die Ziehung der Stichprobe erfolgt aus diesen Schichten, sodass die Stichprobeneinheiten einer Schicht unabhängig von einer anderen Schicht gezogen werden. Im Campus File stehen als Schichtungsinformationen die Variablen Bundesland (ef1) und Gebäudegrößenklasse (ef712) zur Verfügung. Die Regionalschichten sind aus Datenschutzgründen nicht identifizierbar.
- **Endlichkeitskorrektur** (finite population correction; **FPC**): Mit dem Korrekturfaktor für endliche Populationen wird das Ziehen ohne Zurücklegen berücksichtigt. Werte bis Eins werden als schichtspezifischer Auswahlsatz $f_h = n_h / N_h$ interpretiert, Werte darüber als Schichtumfang N_h. Aufgrund des großen Umfangs der Gesamtheit (N) bzw. des geringen Auswahlsatzes des Mikrozensus von einem Prozent der Auswahlbezirke (n/N) liegt FPC ($(N-n) / (N-1)$) sehr nahe bei Eins und ist für die Analyse des Campus Files und des SUF nahezu verzichtbar.
- **Gebundene Hochrechnung** (**Poststratification**): Für den Mikrozensus bis 2004 können die Anpassungsschichten, mit denen die Stichprobendaten nachträglich an bekannte Populationsverteilungen für bestimmte Merkmale angepasst wurden (s. o.), mit der Anweisung `poststrata(varname)` nur auf der regionalen Ebene der Bundesländer definiert werden. Die regionale Anpassungsschicht ist aus Datenschutzgründen nicht identifizierbar.

  Mit `postweight(varname)` werden die den Anpassungsgruppen zugehörigen Populationswerte übergeben, die zuvor mit den Gewichtungsvariablen zur gebundenen Hochrechnung berechnet werden müssen.

Aus didaktischen Gründen werden anhand der folgenden ersten Beispiele die Punkte Designeffekt (4.2) und Gruppenvergleich (4.3) aufgegriffen, die für alle Schätzungen wichtig sind. Dies gilt auch für das Kommando `svydescribe` (Abschn. 4.1), mit dem evtl. bei der Schätzung auftretende Problemfälle geklärt werden können.

4 Gesamtwerte

4.1 Designbasierte Schätzung

Wie in Tabelle 2 zusammenfassend dargestellt, entsprechen die anonymisierten Mikrozensusdaten jeweils einer systematischen Zufallsauswahl aus dem Original-Mikrozensus. Für diesen Fall weisen Varianzschätzungen i. d. R. Verzerrungen auf (Krug et al. 2001: 93f.; Särndal et al. 1997: 73f.). Versuche, die exakten Ziehungswahrscheinlichkeiten zu berücksichtigen, ergeben komplizierte Schätzungen (siehe Gabler und Stenger 2006). In der Praxis müssen deshalb vereinfachende Annahmen getroffen werden.

Vernachlässigt man die geringe Zahl von Haushaltsausfällen, kann die Ziehung der Substichproben Campus File und Scientific Use File als zweiphasiges Ziehungsverfahren (1. Ziehung der realisierten Mikrozensus-Haushalte (1 %); 2. Ziehung der Haushalts- bzw. Wohnungssubstichprobe (Campus File: 3,5 %; Scientific Use File: 70 %) betrachtet werden. Näherungsweise kann auch von einer zweistufigen Auswahl ausgegangen werden (Rendtel und Schimpl-Neimanns 2001: 92; zu den Voraussetzungen siehe Särndal et al. 1997: 133-135). In diesem Sinne entspricht die erste Stufe einer geschichteten Auswahl der PSUs im Original-Mikrozensus. Die zweite Stufe stellt dann die gezogene Substichprobe von Haushalten bzw. Wohnungen in den Erhebungsjahren 1998, 2002 und 2006 usw. dar. Aufgrund des geringen Auswahlsatzes der Auswahlbezirke des Mikrozensus ist anzunehmen, dass die Varianzanteile nach der ersten Stufe vernachlässigbar sind (siehe Särndal et al. 1997: 140).

Der Einfachheit halber können deshalb die anonymisierten Daten – nach Berücksichtigung der Inklusionswahrscheinlichkeit der Substichprobe – wie der Original-Mikrozensus als eine mehrfach geschichtete einstufige Klumpenauswahl betrachtet werden. Diese Annahme ist für das Scientific Use File mit rund 45.000 Auswahlbezirken intuitiv einsichtig, denn aufgrund des hohen Substichprobenauswahlsatzes von 70 Prozent kann man davon ausgehen, dass auf der ersten Auswahlstufe wie im Original-Mikrozensus etwa ein Prozent der Auswahlbezirke der Grundgesamtheit enthalten sind.

Das Campus File umfasst jedoch nur 10.700 Auswahlbezirke mit durchschnittlich 1,1 Haushalten (s. Tab. 1), sodass ein Vorgehen analog zum SUF kaum gerechtfertigt erscheint. Zur Plausibilität verschiedener Annahmen werden nach einer einführenden Übung weitere Testauswertungen durchgeführt und die Ergebnisse abschließend verglichen.

Das Statistische Bundesamt (2003: 19; vgl. auch Krug et al. (2001: 326)) verwendet für die Berechnung des relativen Standardfehlers (Variationskoeffizient, *cv*) eines Gesamtwertes (Total)

(1) $\hat{n}_g = n_g / f$

bei der Designgewichtung bzw. freien Hochrechnung die Formel

(2) $\hat{v}_g^2 = \frac{1-f}{n_g^2} \sum_{h=1}^{L} m_h \cdot s_{gh}^2$

Um die Schätzung der Varianz des Gesamtwertes zu erhalten, muss (2) noch mit $\left(\hat{n}_g^2 / f^2\right)$ multipliziert werden.[8] Die Variablennamen des Stata-Programms (siehe Abschnitt 4.1.1) lehnen sich an diese Notation an und sind in eckigen Klammern genannt.

f	Auswahlsatz, Inklusionswahrscheinlichkeit [f]
L	Anzahl der Schichten
$n_g = \sum_{h=1}^{L} \sum_{i=1}^{m_h} n_{ghi}$	Anzahl der Stichprobenfälle der Merkmalskategorie g [n_g]
n_{ghi}	Anzahl der Stichprobenfälle der Merkmalskategorie *g* im Auswahlbezirk *i* der Schicht *h* [n_ghi]
m_h	Anzahl der Auswahlbezirke in Schicht *h* der Stichprobe [m_h]
$s_{gh}^2 = \frac{1}{m_h - 1} \sum_{i=1}^{m_h} \left(n_{ghi} - \bar{n}_{gh}\right)^2$	Varianz der Stichprobenfälle je Auswahlbezirk in Schicht *h* [s2_gh]
$\bar{n}_{gh} = \frac{1}{m_h} \sum_{i=1}^{m_h} n_{ghi}$	Mittelwert der Stichprobenfälle je Auswahlbezirk in der Schicht *h* und Merkmalskategorie *g* [n_quer_gh]

Die obigen Formeln gehen vom Variationskoeffizienten aus und beziehen sich speziell auf den Mikrozensus. Allgemeiner wird im Stata Survey Data Reference Manual (2007a: 151-152) die einstufige Schätzung eines Gesamtwertes (Total) $\hat{Y}$ und der Varianz $\hat{V}\left(\hat{Y}\right)$ dargestellt. Da Stata für die Schätzungen verwendet wird, werden sie im Folgenden gezeigt. Die der

[8] Implizit angenommen wird hierbei, dass die auf Basis der vorliegenden Stichprobe geschätzte Varianz als ausreichender Ersatz für die unbekannte Populationsvarianz dienen kann.

Notation angelehnten Variablennamen zum Stata-Programm (siehe Abschnitt 4.1.1) sind wieder in eckigen Klammern zu finden.

(1) $$\hat{Y} = \sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} y_{hij}$$

(2) $$\hat{V}\left(\hat{Y}\right) = \sum_{h=1}^{L} (1 - f_h) \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2$$

mit $y_{hi} = \sum_{j=1}^{m_{hi}} w_{hij} y_{hij}$ gewichtetes Total der PSU (h, i) [y_hi]

$\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}$ Mittelwert der PSU-Totals der Schicht h [y_quer_h]

h Schichten h = 1, ..., L

i i-te Primäreinheit (PSU) in der Schicht h mit i = 1, ..., N_h

f_h Auswahlsatz in Schicht h, Inklusionswahrscheinlichkeit [f_h]

N_h Anzahl der PSUs in der Schicht h

n_h Anzahl der PSUs in der Stichprobe in Schicht h [n_h]

m_{hi} Anzahl der Personen in einer PSU i der Schicht h in der Stichprobe

Der mit $w_{hij} = N_h / n_h$ gewichtete Merkmalswert y_{hij} bezieht sich somit auf die Person j im Auswahlbezirk (PSU) i in Schicht h.

Bei Annahme einer zweistufigen Auswahl mit m_{hij} Sekundäreinheiten (SSU) pro PSU in der Stichprobe (z. B. Haushalten) werden der designgewichtete Gesamtwert und die Varianz in Stata (2007a: 152-154) analog zur einstufigen Auswahl wie folgt geschätzt:

$$\hat{Y} = \sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \sum_{k=1}^{m_{hij}} w_{hijk} y_{hijk}$$

Hier steht y_{hijk} für das Merkmal der Person k in Haushalt j usw. Das Designgewicht ist das Produkt der Kehrwerte der Auswahlsätze auf der ersten und zweiten Stufe: $w_{hijk} = \left(N_h / n_h\right) \cdot \left(M_{hi} / m_{hi}\right)$.

Die geschätzte Varianz ist:

$$\hat{V}\left(\hat{Y}\right) = \sum_{h=1}^{L}(1-f_h)\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}(y_{hi}-\bar{y}_h)^2 + \sum_{h=1}^{L}f_h\sum_{i=1}^{n_h}(1-f_{hi})\frac{m_{hi}}{m_{hi}-1}\sum_{j=1}^{m_{hi}}(y_{hij}-\bar{y}_{hi})^2$$

wobei y_{hi} und y_{hij} die jeweils mit w_{hijk} gewichteten y-Werte bezeichnen.

Der Ausdruck in der zweiten Zeile bildet die durch die Substichprobenauswahl zusätzlich entstehende Varianz ab (vgl. Rendtel und Schimpl-Neimanns 2001). Der Auswahlsatz der ersten Stufe der Auswahlbezirke ist f_h, f_{hi} ist der Auswahlsatz der zweiten Auswahlstufe.

Mit **`svyset`** wird das Stichprobendesign des Campus Files nach Einlesen der Daten und Bildung der benötigten Variablen definiert. Als Beispiel soll zunächst unter der Annahme einer einstufigen Ziehung die Zahl von Erwerbslosen (Gesamtwert bzw. Total) geschätzt werden. Das Gewicht w_{hij} entspricht der mit `svyset (...) [pw = w]` angegebenen Gewichtungsvariable. Die interessierende Subpopulation ist die Bevölkerung am Hauptwohnsitz im (erwerbsfähigen) Alter von 15 bis 65 Jahren. Für die Definition der Schichten werden zunächst nur die in den anonymisierten Daten vorliegenden Schichtungsinformationen Bundesland (ef1) und Gebäudegrößenklasse (ef712) herangezogen.

An dieser Stelle ist darauf aufmerksam zu machen, dass bei der Varianzschätzung die nicht interessierenden Einheiten keinesfalls durch eine Fallselektion (z. B. mit „if“ oder „keep if“) ausgeschlossen werden dürfen, da sonst die Berechnung des Standardfehlers fehlerhaft wird (Graubard und Korn 1996). Stattdessen muss die Option „`subpop(variable)`“ verwendet werden, mit der eine Indikatorvariable (1 = interessierende Subpopulation, 0 = sonst) benannt wird. Falls nach verschiedenen Subpopulationen (z. B. Geschlecht) differenziert werden soll, steht noch die Option „`over(variable)`“ zur Verfügung.

Für die Schätzung unter **Annahme einer einstufigen Klumpenauswahl** wird die Ziehungswahrscheinlichkeit eines Auswahlbezirks im Campus File benötigt. In den folgenden ersten Auswertungen wird stark vereinfachend davon ausgegangen, dass diese Ziehungswahrscheinlichkeit dem Produkt der Ziehungswahrscheinlichkeiten eines Auswahlbezirks im Original-Mikrozensus (1 %) und der Substichprobe Campus File (3,5 %) entspricht und damit die Zie-

hungswahrscheinlichkeit eines Auswahlbezirks gleich der Ziehungswahrscheinlichkeit von Wohnungen, Haushalten und Personen ist.

Beispiel 1: Designbasierte Schätzung des Gesamtwertes Zahl der Erwerbslosen
a) Ohne Ausschluss von PSUs mit nur einer Sekundäreinheit
– Annahme einer einstufigen Zufallsauswahl (f=0,01*0,035)–

```
use ef1 ef3 ef30 ef504 ef505 ef506 ef712 using mz02cf_Beisp.dta, replace

* Schichtung: Bundesland (ef1), Gebäudegrößenklasse (ef712)
gen schicht = ef1*10 + ef712
gen f = (0.01 * 0.035) // Auswahlsatz MZ: 1%, CF: 3,5%
gen w = 1/f            // Designgewicht

* Definition des Stichprobendesigns
* - Einstufige Klumpenauswahl: Auswahlbezirke (ef3)
* - Schichtung: Bundesland (ef1), Gebäudeschicht (ef712)
* - Endlichkeitskorrektur mit Auswahlsatz f = 0,00035
svyset ef3 [pw=w] , strata(schicht) fpc(f)

* Konstruktion der interessierenden Variablen
gen y = ef504==2 // ILO-Erwerbslose

* sub: Subpopulation: Bevölkerung am Hauptwohnsitz, 15+ Jahre
gen sub = ef505>=1 & ef505<=2 & ef30>=15

* Schätzung des Gesamtwertes
svy linearized, subpop(sub) : total y

  Survey: Total estimation
  Number of strata =      79      Number of obs    =     25137
  Number of PSUs   =   10707      Population size  =  71820005
                                  Subpop. no. obs  =     21193
                                  Subpop. size     =  60551433
                                  Design df        =     10628

             |             Linearized
             |      Total   Std. Err.   [95% Conf. Interval]
-------------+------------------------------------------------
           y |    2874286          .           .           .
-------------------------------------------------------------
  Note: missing standard error because of stratum with single sampling
  unit.
```

Das Ergebnis zeigt, dass die Population mit Designgewichtung nur auf rund 72 Millionen geschätzt wird (siehe „Population size“). Es wird lediglich der Gesamtwert der Erwerbslosen ($\hat{t} = 2.874.286$), aber kein Standardfehler berechnet. Es gibt offensichtlich eine Schicht, die nur einen Auswahlbezirk umfasst, sodass keine Varianzschätzung möglich ist.

In dem obigen Beispiel wurde die voreingestellte Option `singleunit(missing)` angewendet. Mit den Optionen `singleunit(certainty)` können die Varianzen solcher Schichten bei der Berechnung des Standardfehlers auf Null gesetzt und somit ausgeschlossen werden. Für die Imputation von Schicht- oder Gesamt-Durchschnittswerten stehen weitere Optionen zur Verfügung (`singleunit(scaled)` und `singleunit(centered)`).

Setzt man die betreffenden Einheiten bzw. Schichten bei der Varianzschätzung auf Null, erhält man:

Beispiel 1: Designbasierte Schätzung des Gesamtwertes Zahl der Erwerbslosen
b) Ausschluss von PSUs mit nur einer Sekundäreinheit

```
svyset ef3 [pw=w] , strata(schicht) fpc(f) singleunit(certainty)
svy linearized, subpop(sub) : total y
```

```
                |             Linearized
                |      Total   Std. Err.   [95% Conf. Interval]
  --------------+-----------------------------------------------
              y |    2874286   94763,38     2688532     3060040
  ------------------------------------------------------------
  Note: strata with single sampling unit treated as certainty units.
```

Um mögliche Verzerrungen bei diesem Vorgehen abschätzen zu können, ist es aber ratsam, die Problemfälle mit dem Kommando `svydescribe` zu untersuchen. In diesem Fall betrifft es nur eine nicht erwerbstätige Person, die zur Bevölkerung in Gemeinschaftsunterkünften (ef506=2) gehört, sodass keine Auswirkungen auf die Varianzschätzung zu erwarten sind.

```
svydescribe, single gen(single)
{...}
list ef1 ef712 ef3 ef505 ef506 y sub if single, nolab noobs
```

```
     | ef1   ef712    ef3   ef505   ef506   y   sub |
     |-------------------------------------------------|
     |   4       4   1707       1       3   0     1 |
     +-------------------------------------------------+
```

Ansonsten könnte dieser Fall aus Bremen (ef1=4) auch mit ähnlichen Gruppen (z. B. Gemeinschaftsunterkünfte (ef712) in Niedersachsen (ef1=3)) zu einer sogenannten Pseudoschicht zusammengefasst werden. Das folgende Beispiel 1c) zeigt, dass der gleiche Standardfehler wie beim Ausschluss aus der Varianzschätzung geschätzt wird.

Beispiel 1: Designbasierte Schätzung des Gesamtwertes Zahl der Erwerbslosen
c) Rekodierung von PSUs mit nur einer Sekundäreinheit (Pseudoschicht)

```
* Pseudoschicht: Bremen & Gemeinschaftsunterk. => Niedersachsen
gen pschicht = schicht
replace pschicht = 34 if single
svyset ef3 [pw=w] , strata(pschicht) fpc(f) singleunit(missing)
svy linearized, subpop(sub) : total y
```

```
                |             Linearized
                |      Total   Std. Err.   [95% Conf. Interval]
  --------------+-----------------------------------------------
              y |    2874286   94763,38     2688532     3060040
  ------------------------------------------------------------
```

Diese Schätzung beruht auf der Annahme (A) einer identischen Ziehungswahrscheinlichkeit von Auswahlbezirken, Wohnungen, Haushalten und Personen im Campus File (f_i = 0,00035).

Alternativ kann man (B) das Campus File als eine einfache Zufallsauswahl von Wohnungen betrachten und die Wohnung als Primäreinheit definieren. Da hierbei jedoch ignoriert wird, dass immerhin rund zehn Prozent der Auswahlbezirke des Campus File mehr als eine Wohnung (bzw. mehr als einen Haushalt) umfassen, dürfte das mit einer Unterschätzung der Varianz verbunden sein.

Die obige Annahme (A) der Ziehungswahrscheinlichkeit (f_i = 0,00035) erscheint allerdings wenig plausibel, wie der Vergleich der Anzahl der Auswahlbezirke im Campus File und Scientific Use File zeigt (s. Tab. 1), denn mit dieser Annahme wird die Zahl der Auswahlbezirke im Original-Mikrozensus auf 305.914 (= 10.707 / 0,035) geschätzt. Geht man davon aus, dass die Zahl der Auswahlbezirke des SUF (n = 45.058) sehr nahe bei der des Original-Mikrozensus liegt, lässt sich die Ziehungswahrscheinlichkeit der ersten Stufe im Campus File auf etwa 0,24 Prozent schätzen: f_i = 0,01 * 10.707 / 45.058 = 0,002376.

Aus diesem Grund ist bei einer **zweistufigen Schätzung** im Campus File zu beachten, dass sehr viele Schichten nur eine Einheit der zweiten Stufe enthalten. Testhalber wird in Version (C) als Ziehungswahrscheinlichkeit der ersten Stufe von Auswahlbezirken f_i = 0,002376 angenommen. Die bedingte Wahrscheinlichkeit für die Ziehung eines Haushalts bzw. einer Wohnung *j* im Campus File aus einem Auswahlbezirk *i* der Gesamtheit beträgt dann rund 0,15 Prozent ($f_{j|i}$ = 0,01 * 0,035 / 0,00237627 = 0,1473). Das Designgewicht entspricht dem Kehrwert der Inklusionswahrscheinlichkeit der Sekundäreinheiten (Wohnungen, Haushalte oder Personen) und zugleich dem Kehrwert des Produkts der beiden (bedingten) Wahrscheinlichkeiten ($w_{ij} = 1/\left(f_i \cdot f_{j|i}\right)$.[9]

Schließlich werden in Version (D) als Ziehungswahrscheinlichkeiten der PSUs die des Original-Mikrozensus (1 %) und für die zweite Stufe von Haushalten die des Campus Files (3,5 %) angenommen, auch wenn sie nicht überzeugend sind.[10]

Für die in den anonymisierten Daten nicht identifizierbare Regionalschicht kann ggf. die Variable Gemeindegrößenklasse (ef708) als Proxy-Variable bei der Definition der Schichtung (siehe Variable „schicht2“) verwendet werden. Da sich dadurch die Schichtumfänge reduzie-

[9] Da die Wohnungssubstichprobe des Campus File alle Personen einer ausgewählten Sekundäreinheit umfasst, ist w_{hijk} gleich mit w_{ij}.

[10] Es ergeben sich in diesem Beispiel keine Unterschiede, wenn für die zweite Stufe Wohnungen statt Haushalte definiert werden.

ren, treten mehr singuläre PSUs pro Schicht auf.[11] Die obigen Versionen werden ebenfalls mit dieser erweiterten Schicht als A'-D' geschätzt.

Tabelle 3 zeigt, wie sich die jeweiligen Annahmen in unterschiedlichen Schätzungen des Stichprobenfehlers niederschlagen. Betrachtet man zunächst die mit den Variablen Bundesland (ef1) und Gebäudegrößenklasse (ef712) definierte Schichtung, ist, wie zu erwarten, bei den weniger realistischen Versionen B und D die Standardabweichung am geringsten. Die vermutlich plausibelsten Schätzungen ergeben sich mittels der nachträglich für das Campus File berechneten Inklusionswahrscheinlichkeit von Auswahlbezirken. Dabei liegen die ein- und zweistufigen Schätzungen A und C nahe zusammen. Dies belegt, dass die vereinfachte Annahme einer einstufigen Auswahl kaum Einfluss auf die Varianzschätzung hat. Der größte Standardfehler wird allerdings mit Version A geschätzt, die von einer einstufigen Auswahl von PSUs ausgeht. Durch die ergänzend mit der Gemeindegrößenklasse (ef708) differenziertere Schichtung reduzieren sich die geschätzten Standardabweichungen.

Tabelle 3: Ergebnisse der Varianzschätzungen zum designgewichteten Gesamtwert der Zahl der Erwerbslosen ($\hat{t} = 2.874.286$) unter verschiedenen Varianten (Beispiele 1b und 1d)

			Schicht: ef1, ef712		Schicht2: + ef708	
	1. Stufe; Auswahlsatz	2. Stufe; Auswahlsatz	Standard-abw.	CV (%)	Standard-abw.	CV (%)
A	PSU; f = f_{MZ} * f_{CF} = 0,00035		94.763,38	3,30	94.660,37	3,29
B	Wohnung; f = 0,00035		94.093,84	3,27	93.995,55	3,27
C	PSU; f1 = 0,0024	Haushalt; f2 = 0,1473	94.684,76	3,29	94.581,88	3,29
D	PSU; f_{MZ} = 0,01	Haushalt; f_{CF} = 0,035	94.388,37	3,28	94.285,90	3,28

Geht man davon aus, dass die Ergebnisse dieses Beispiels auch auf andere Merkmale übertragen werden können, sind somit sowohl zweistufige Schätzungen als auch die erweiterte Schichtung einsetzbar. Die Variationskoeffizienten betragen jeweils etwa 3,3 Prozent. Die Unterschiede sind insgesamt betrachtet marginal. Da mit Version A somit ohne großen Aufwand eine konservative Schätzung des Stichprobenfehlers im Campus File möglich scheint,

[11] Dies betrifft 15 Schichten bzw. PSUs der ersten Stufe mit insgesamt 17 Personen unter denen elf Personen der Bevölkerung in Gemeinschaftsunterkünften sind. Testauswertungen zeigten, dass sich nach einer Rekodierung der Problemfälle zu Pseudoschichten die Standardabweichungen gegenüber der Behandlung mit der Option `singleunit(certainty)` nur unwesentlich verändern. Der Einfachheit halber werden deshalb die betreffenden PSUs bei der Varianzschätzung mit dieser Option auf Null gesetzt.

dürfte diese Annahme – ohne weitergehende Ansprüche – für die folgenden exemplarischen Übungen tragbar sein.

Beispiel 1: Designbasierte Schätzung des Gesamtwertes Zahl der Erwerbslosen
d) Alternative Schätzungen

```
{...}
* (B) Einstufige Schätzung
*     PSU=Wohnung, Auswahlsatz f = 0,00035
gen whg = ef3*10+ef7
replace w = 1/f
svyset whg [pw = w], strata(schicht) fpc(f) single(certainty)
svy linearized, subpop(sub) : total y

* (C) Zweistufige Schätzung,
*     1. PSU=Auswahlbezirk, Auswahlsatz f1 = 0,002376
*     2. SSU=Haushalt, Auswahlsatz      f2 = 0,00035 / f1 = 0,147
gen f2 = f / f1
replace w = 1/(f1*f2)
svyset ef3 [pw=w], strata(schicht) fpc(f1) single(certainty) ///
    || ef4, fpc(f2)
svydescribe,  stage(2) single  gen(s2)
tab s2 s1
svy linearized, subpop(sub) : total y

* (D) Zweistufige Schätzung,
*     1. PSU=Auswahlbezirk, Auswahlsatz f_MZ = 0,01
*     2. SSU=Haushalt, Auswahlsatz f_CF = 0,035
gen f_MZ = 0.01
gen f_CF = 0.035
replace w = 1/(f_MZ * f_CF)
svyset ef3 [pw=w], strata(schicht) fpc(f_MZ) single(certainty) ///
    || ef4, fpc(f_CF)
svy linearized, subpop(sub) : total y
* === Erweiterte Schichtung: + Gemeindegrößenklasse ===
gen schicht2 =  ef1*100 + ef708*10 + ef712

* (A') Einstufige Schätzung, schicht: + Gemeindegrößenklasse
replace w = 1/f
svyset ef3 [pw = w], strata(schicht2) fpc(f) single(certainty)
svydescribe, stage(1) single gen(s_A)
svy linearized, subpop(sub) : total y
{...}

* (D') Zweistufige Schätzung, schicht: + Gemeindegrößenklasse
replace w = 1/(f_MZ * f_CF)
svyset ef3 [pw=w], strata(schicht2) fpc(f_MZ) vce(linearized) ///
   singleunit(certainty) || ef4, fpc(f_CF)
svydescribe, stage(1) single gen(s_E1)
svydescribe, stage(2) single gen(s_E2)
svy linearized, subpop(sub) : total y
```

4.1.1 Ein Berechnungsbeispiel

Die hier bei der Schätzung verwendete Annahme (A) einer mehrfach geschichteten einstufigen Klumpenauswahl entspricht im Wesentlichen der Annahme, die den Fehlerrechnungen

des Statistischen Bundesamtes für den Original-Mikrozensus zugrunde liegt. Mit einem einfachen Beispiel wird gezeigt, wie mit den Formeln des Statistischen Bundesamtes sowie den Stata-Formeln die Schätzungen „von Hand“ durchgeführt werden können. Der Übersichtlichkeit halber wird nur eine kleine Teilauswahl von zehn Auswahlbezirken betrachtet und keine Schichtung berücksichtigt.

Beispiel 2: Designbasierte Schätzung der Zahl der Erwerbstätigen (10 PSUs)

```
use ef1 ef3 ef4 ef30 ef504 ef505 ef506 ef712 ///
    if ef3==0187 | ef3==2353 | ef3==3084 | ef3==4353 | ///
       ef3==6555 | ef3==6555 | ef3==6579 | ef3==7825 | ///
       ef3==8517 | ef3==9330 | ef3==9909 using mz02cf_Beisp.dta, replace
gen f = 0.01 * 0.035 // Auswahlsatz MZ 1%, CF 3,5%
gen w = 1/f          // Designgewicht
* Subpopulation Bev. in Privathaushalten am Hauptwohnsitz
*               im Alter von 15 Jahren und älter
gen sub = ef506==1 & ef505>=1 & ef505<=2 & ef30>=15
gen y = ef504==1 // ILO-Erwerbstätige ggf. mit sub multipliz.

* Individualdaten
list ef3 y sub w, nolab sepby(ef3)
```

	ef3	y	sub	w
1.	187	0	1	2857,143
2.	2353	0	1	2857,143
3.	3084	1	1	2857,143
4.	3084	1	1	2857,143
5.	4353	1	1	2857,143
6.	4353	0	1	2857,143
7.	4353	1	1	2857,143
8.	6555	0	1	2857,143
9.	6579	1	1	2857,143
10.	6579	1	1	2857,143
11.	6579	0	0	2857,143
12.	6579	0	0	2857,143
13.	6579	0	0	2857,143
14.	7825	1	1	2857,143
15.	7825	1	1	2857,143
16.	8517	1	1	2857,143
17.	8517	0	1	2857,143
18.	9330	1	1	2857,143
19.	9330	1	1	2857,143
20.	9330	0	0	2857,143
21.	9909	0	1	2857,143

```
* Einstufige Klumpenausw., ohne Schichtung, Designgewichtung
* - FPC nicht korrekt; nur für dieses Beispiel
svyset ef3 [pw = w], single(missing) fpc(f)
svy linearized, subpop(sub) : total y
{...}
                |             Linearized
                |      Total   Std. Err.   [95% Conf. Interval]
   -------------+----------------------------------------------
              y |   31428,57   8983,172    11107,23    51749,92
   ------------------------------------------------------------
```

Unter der Annahme, dass die Sekundäreinheiten (Haushalte, Personen) eines Auswahlbezirks (PSU) vollständig erfasst werden, kann die Schätzung des Stichprobenfehlers auf der PSU-Ebene durchgeführt werden. Die Aggregation erfolgt mittels `collapse`.

```
/* Berechnen der ungewichteten PSU Totals für Anwendung der StBA-Formeln
   sowie der gewichteten PSU Totals und Anwendung der Stata-Formel */

collapse (sum)  n_ghi=y , by(ef3)
gen y_hi = n_ghi * 1/(0.01 * 0.035)
list , nolab sep(0)
```

	ef3	n_ghi	y_hi
1.	187	0	0
2.	2353	0	0
3.	3084	2	5714,286
4.	4353	2	5714,286
5.	6555	0	0
6.	6579	2	5714,286
7.	7825	2	5714,286
8.	8517	1	2857,143
9.	9330	2	5714,286
10.	9909	0	0

```
* ==== Umsetzung StBA-Formeln ====
scalar f = 0.01 * 0.035  // Auswahlsatz
scalar m_h = _N
sum n_ghi, meanonly
scalar n_quer_gh = 1/_N * r(sum)
egen s2_gh = total((n_ghi - n_quer_gh)^2)
replace s2_gh = 1/(m_h-1) * s2_gh
scalar s2_gh = s2_gh[_N]
* (2) mit (n^2_g / f^2) multiplizieren
scalar var_y = (1-f) * m_h * s2_gh / (f*f) /* Varianz */
scalar s_y = var_y^.5                     /* Std.fehler */
* (1) mit Designgewichtung
gen total_y = sum(n_ghi*1/f)              /* Gesamtwert */
display as text "Total:  ", as res total_y[_N], ///
        _newline(1) as text "Std.Abw.:", as res s_y, ///
        _newline(1) as text "CV (%):    ", as res s_y*100/total_y[_N]
```

```
  Total:   31428,57
  Std.Abw.: 8983,17
  CV (%):     28,58
```

```
* ==== Umsetzung Stata-Formeln ====
scalar f_h = 0.01 * 0.035  // Auswahlsatz
```

```
scalar n_h = _N
sum y_hi, meanonly
scalar y_quer_h = 1/_N * r(sum)
egen v_y = total((y_hi - y_quer_h)^2)
replace v_y = (1-f_h) * n_h/(n_h-1) * v_y  /* Varianz */
scalar se_y = v_y[_N]^.5                   /* Std.fehler */
gen t_y = sum(y_hi)                        /* Gesamtwert */
display as text "Total:  ", as res t_y[_N], ///
        _newline(1) as text "Std.Abw.:", as res se_y, ///
        _newline(1) as text "CV (%):    ", as res se_y*100/t_y[_N]
```

```
  Total:   31428,57
  Std.Abw.: 8983,17
  CV (%):     28,58
```

Wie zu sehen ist, sind die obigen Stata-Ergebnisse mit den jeweiligen Formeln einfach „von Hand“ zu replizieren.

4.2 Designeffekte

Die im Mikrozensus befragten Haushalte und Personen ausgewählter Auswahlbezirks benachbarter Wohnungen sind sich in vielen Merkmalen ähnlicher, als wenn eine gleiche Zahl von Haushalten und Personen durch eine einfache Zufallsstichprobe ausgewählt worden wären. Tendenziell vergrößert die durch die Ziehung von Auswahlbezirken als Primäreinheiten bezeichnete Klumpung die Varianz der Populationsschätzer im Vergleich zu einer einfachen Zufallsauswahl. Dagegen ist eine geschichtete Auswahl mit einer Verringerung der Varianz verbunden, wenn im Idealfall die Merkmale innerhalb der Schichten möglichst homogen und zwischen den Schichten heterogen verteilt sind.

Der Designeffekt beschreibt das Verhältnis des Standardfehlers, der unter Berücksichtigung der Annahmen des Stichprobendesigns geschätzt wurde, zum Standardfehler einer hypothetischen Stichprobenziehung gleichen Umfangs, die jedoch ohne Klumpung und Schichtung durchgeführt worden wäre – also unter der Annahme einer einfachen Zufallsstichprobe, die üblicherweise den Standardprogrammen zugrunde liegt.

Das Statistische Bundesamt veröffentlicht Designeffekt-Faktoren für ausgewählte Merkmale und Subpopulationen (z. B. Statistisches Bundesamt 2003: 20ff.). Diese als „Zuschlagsfaktoren“ bezeichneten Designeffekte können jedoch bei der Analyse des Campus Files oder des Scientific Use Files nur eingeschränkt verwendet werden, da sich durch die Substichprobenziehung die Zahl der Haushalte und Personen pro Auswahlbezirk im Vergleich zum Original-Mikrozensus und damit auch die Klumpeneffekte verringern (siehe Rendtel und Schimpl-Neimanns 2001).

Mithilfe der in den Daten enthalten Informationen über die Klumpung und Schichtung können Designeffekte mit Stata ermittelt werden. Die Ausgabe von Designeffekten wird mit `estat effects` nach der Schätzung von Statistiken (z. B. nach `svy (...): total (...)`) angefordert. Stata unterscheidet zwei Designeffektfaktoren: *DEFF* und *DEFT*. Zusätzlich können noch Kennziffern zur Fehlspezifikation (**m**isspecification **eff**ects) *MEFF* und *MEFT* ausgegeben werden, die bei der Beurteilung von statistischen Modellen eine größere Rolle spielen als bei der Populationsschätzung.[12]

DEFF gibt das Verhältnis der designbasierten Schätzung der Varianz eines Parameters $\hat{V}(\theta)$ zu einer Schätzung unter Annahme einer einfachen Zufallsstichprobe $\hat{V}_{srswor}(\theta_{srs})$ wieder (*srswor*: Ziehen ohne Zurücklegen). Im Unterschied dazu bezieht sich *DEFT* auf den designbasierten Standardfehler im Verhältnis zum Standardfehler, der unter Annahme einer einfachen Zufallsstichprobe mit Zurücklegen (*srswr*) geschätzt wurde (Stata 2007a: 43-45). Da für die Mikrozensusdaten die Endlichkeitskorrektur für das Ziehen ohne Zurücklegen unerheblich ist, unterscheiden sich $DEFF^{1/2}$ und *DEFT* praktisch nicht.

$$DEFF = \frac{\hat{V}(\theta)}{\hat{V}_{srswor}(\theta_{srs})}$$

$$DEFT = \sqrt{\frac{\hat{V}(\theta)}{\hat{V}_{srswr}(\theta_{srs})}}$$

Bei den Fehlspezifikationseffekten *MEFF* bzw. *MEFT* wird die designbasierte Varianz bzw. Standardabweichung im Verhältnis zu einer Schätzung unter der Annahme einer einfachen Zufallsauswahl mit Zurücklegen (srswr) betrachtet, bei der Gewichtung, Schichtung und Klumpung fälschlicherweise ignoriert werden.

$$MEFF = \frac{\hat{V}(\theta)}{\hat{V}_{msp}(\theta_{msp})} \qquad MEFT = \sqrt{MEFF}$$

Im Beispiel 1 (A) beträgt der Standardfehler des Gesamtwertes der Zahl der Erwerbslosen 94.763. Nach Anforderung der Designeffekte sieht man, dass der designbasierte Standardfehler rund sieben Prozent höher ist als bei einer Schätzung, wie sie üblicherweise mit den Standardformeln berechnet wird (siehe *DEFT* und *MEFT*). Dementsprechend ist das Konfidenzin-

[12] Die Schreibweise von *DEFF*, *DEFT*, *MEFF* und *MEFT* entspricht dem Stata Manual. Es ist aber zu beachten, dass es sich bei diesen Effekten um Schätzungen auf Basis der Stichprobe handelt.

tervall bei designbasierter Schätzung breiter. Im folgenden Beispielprogramm wird ergänzend gezeigt, wie man die Effekte selbst berechnen kann.

Beispiel 3a: Designeffekte im Campus File

```
* Definitionen wie im Beispiel 1b, y = Erwerbslos
(...)
svyset ef3 [pw=w] , strata(schicht) fpc(f) singleunit(certainty)
svy linearized, subpop(sub) : total y
estat effects, deff deft meff meft
```

```
       |          Linearized
       |   Total  Std. Err.     DEFF     DEFT    MEFF      MEFT
  -----+-------------------------------------------------------
     y | 2874286  94763,38   1,13944  1,06726  1,14794  1,07142
```

```
matrix v_d = e(V)   /* Designbasierte Varianz */
svmat  v_d

* SRSWOR: Einf. Zufallsstichprobe, Ziehen ohne Zurücklegen
svyset _n [pw=w], fpc(f)
svy linearized, subpop(sub) : total y
matrix v_srswor = e(V)
svmat  v_srswor

* SRSWR: Einfache Zufallsstichprobe, Ziehen mit Zurücklegen
svyset _n [pw=w]
svy linearized, subpop(sub) : total y
matrix v_srswr = e(V)
svmat  v_srswr

* SRSWR wie für Berechnung MEFF und MEFT
total y [pw=w] if sub
matrix v_msp = e(V)
svmat  v_msp

gen DEFF = v_d/v_srswor
gen DEFT = (v_d/v_srswr)^.5
gen MEFF = v_d/v_msp
gen MEFT = (v_d/v_msp)^.5
```

```
list DEFF DEFT MEFF MEFT in 1/1, noobs
```

```
     |     DEFF       DEFT      MEFF      MEFT |
     |-----------------------------------------|
     | 1,139441   1,067259   1,14794   1,07142 |
```

Im Campus File umfassen die meisten Auswahlbezirke nur einen Haushalt, sodass die damit berechneten Designeffektfaktoren im Wesentlichen die Klumpung auf Haushaltsebene widerspiegeln. Schätzt man für das Scientific Use File des Mikrozensus 2002 das obige Beispiel 3, ergibt sich, abgesehen von dem aufgrund der anderen Stichprobe abweichenden Gesamtwert (Total), ein um rund neun Prozent größerer Designeffekt (*DEFT* SUF: 1,16; CF: 1,07).

Beispiel 3b: Designeffekte im Scientific Use File

```
* Stata-Programm wie Beispiel 1 und Beispiel 3a, Y = Erwerbslos
* im SUF abweichende Stichprobendefinitionen
* gen f = 0.01
* gen w = 1/(0.01 * 0.70)  // Designgewicht
{...}
estat effects, deff deft meff meft
```

	Linearized					
	Total	Std. Err.	DEFF	DEFT	MEFF	MEFT
y	2930572	23156,14	1,34465	1,15552	1,34622	1,16027

4.3 Gruppenvergleiche

In der Praxis ist man oft nicht nur an Gesamt- oder Anteilswerten interessiert, sondern will diese Schätzungen für verschiedene Subgruppen vergleichen. Die SVY-Kommandos zählen zu den sogenannten Präfix-Kommandos, sodass designbasierte Auswertungen auf viele weitere Arten vorgenommen werden können; z. B. zweidimensionale Tabellen oder statistische Modelle. Dies soll am Beispiel des Gesamtwertes Erwerbsloser gezeigt werden, die nach Region (West- vs. Ostdeutschland) gegliedert werden. Die Subpopulation wird im Unterschied zum obigen Beispiel enger definiert und umfasst Erwerbspersonen im Alter von 15 bis 65 Jahren am Hauptwohnsitz. Die letzte Tabelle enthält die Erwerbslosenquoten und zeigt die statistisch signifikant in West (6,1 %) und Ost (18,2 %) verschiedene Arbeitsmarktlage (vgl. Breiholz 2003: 606). Der Unabhängigkeitstest wird mit einer Korrektur zur Berücksichtigung des Stichprobendesigns durchgeführt. Die entsprechende designbasierte Pearson χ^2-Statistik ist um mehr als die Hälfte (= 304,2 / 733,7) niedriger als der unter der nicht zutreffenden Standardannahme unabhängig identisch verteilter (i. i. d.) Beobachtungen geschätzte χ^2-Wert.

Beispiel 4: Gruppenvergleiche

```
* Stata-Programm analog zu Beispiel 1b und 3a
gen f = 0.01 * 0.035
gen w = 1/f

svyset ef3 [pw=w] , strata(schicht) fpc(f) singleunit(certainty)
recode ef504 (2=1 "Erwerbslos") (*=0 "Sonst"), gen(y)
label var y "ILO-Erwerbsstatus" // ! nur bei Abgrenzung mit „sub“
* Subpop.: Bev. am Hauptwohnsitz, 15-65 Jahre, Erwerbspers.
gen sub = ef505<=2 & ef30>=15 & ef30<=65 & ef504<=2
* West-/Ostdeutschland
recode ef1 (1/11=1 "West") (*=2 "Ost"), gen(westost)
replace westost = 2 if ef708==9  /* Ost-Berlin */
label var westost "West-/Ostdeutschland"
```

```
* Schätzung des Gesamtwertes Zahl der Erwerbslosen
svy linearized, subpop(sub) : total y

        |             Linearized
        |      Total   Std. Err.     [95% Conf. Interval]
  ------+------------------------------------------------
      y |    2874286   94763,38       2688532     3060040

* ... in West-/Ostdeutschland
svy linearized, subpop(sub) : total y, over(westost)

             |             Linearized
        Over |      Total   Std. Err.     [95% Conf. Interval]
  -----------+------------------------------------------------
  y          |
        West |    1677143   72162,38       1535691     1818595
         Ost |    1197143   61777,29       1076048     1318238

* ... Fallzahltabelle
svy linearized, subpop(sub) : tabulate y westost, count ///
                format(%8.0f) cellwidth(10) stubwidth(20)

                        |        West-/Ostdeutschland
     ILO-Erwerbsstatus |       West         Ost       Total
  ---------------------+----------------------------------
                 Sonst |   25697145     5394286    31091431
            Erwerbslos |    1677143     1197143     2874286
                       |
                 Total |   27374288     6591429    33965717
{...}

* ... mit Spalten-% und Unabhängigkeitstest
svy linearized, subpop(sub): tabulate y westost, col ///
    deft pearson  format(%5.3f) cellwidth(15) stubwidth(20)

  ------------------------------------------------------------
                      |     West-/Ostdeutschland
   ILO-Erwerbsstatus |        West           Ost         Total
  -------------------+----------------------------------------
               Sonst |       0,939         0,818         0,915
                     |       1,052         1,080         0,322
                     |
          Erwerbslos |       0,061         0,182         0,085
                     |       1,052         1,080         0,322
                     |
               Total |       1,000         1,000         1,000
                     |
  ------------------------------------------------------------
    Key:  column proportions
          deft for variances of column proportions
    Pearson:
      Uncorrected   chi2(1)         =  733,6766
      Design-based  F(1, 10585)     =  304,2021     P = 0,0000
    Mean generalized deff           =    4,2422
    CV of generalized deffs         =    0,0000
  Note: 7 strata omitted because they contain no subpopulation members.
```

4.4 Gebundene Hochrechnung (Poststratifikation)

In den obigen Hochrechnungen wurden lediglich die Stichprobenwerte berücksichtigt. Diese Schätzungen unterscheiden sich jedoch von den veröffentlichen Ergebnissen der statistischen Ämter. Während beispielsweise die Zahl erwerbsloser Personen mit designgewichteter Auswertung auf rund 2,9 Millionen (s. o.) geschätzt wird, berichtet das Statistische Bundesamt rund 3,5 Millionen (Statistisches Bundesamt 2003b: 161; siehe auch Breiholz 2003: 603). Die Differenz ist hauptsächlich darauf zurückzuführen, dass die statistischen Ämter die Fallzahlen des Mikrozensus bei der Hochrechnung an Populationswerte der laufenden Bevölkerungsfortschreibung anpassen. Dieses auch als nachträgliche Schichtung (Poststratifikation, Redressment) bezeichnete Vorgehen soll systematische Unter- und Übererfassungen reduzieren, die z. B. durch Befragungsausfälle oder Mängel bei der Erfassung von Neubauten (Herberger 1985: 35) entstehen können. Allgemein setzt eine wirksame Korrektur von Verzerrungen eine hohe Korrelation zwischen Ausfall bzw. Unter- oder Übererfassung mit den Anpassungsmerkmalen voraus (Krug et al. 2001: 202). Selbst falls dies aufgrund der bei der Anpassung verwendeten groben soziodemografischen Merkmale nicht oder nur ansatzweise zutrifft, kann dadurch zumindest erreicht werden, dass die entsprechend hochgerechneten Ergebnisse des Mikrozensus hinsichtlich der angepassten Populationsmerkmale konsistent sind.

Bei der Poststratifikation erfolgte bis 2004 die Anpassung an Ergebnisse der laufenden Bevölkerungsfortschreibung für die Merkmalskombinationen Geschlecht und Staatsangehörigkeit (Deutsche / Ausländer) auf der regionalen Ebene von sogenannten Anpassungsschichten sowie für Soldaten und Wehrpflichtige an entsprechende Bestandsmeldungen auf Regierungsbezirksebene (siehe Heidenreich 1994; Statistisches Bundesamt 2003a). Insgesamt ergeben sich daraus sechs disjunkte Anpassungsklassen pro regionaler Einheit (Anpassungsschicht bzw. Regierungsbezirk), die in den anonymisierten Daten allerdings nur auf Ebene der Bundesländer identifizierbar sind. Der personenbezogene Anpassungsfaktor wird als Quotient der Eckwerte der Populationsdaten („Soll“) und der nach Korrektur der Haushaltsausfälle gewichteten Mikrozensusdaten („Ist“) pro Anpassungsklasse und Anpassungsschicht ermittelt („Soll durch Ist“). Ab 1990 wurde für Haushalts- und Personenauswertungen ein Haushaltsfaktor eingeführt, der als arithmetisches Mittel der personenbezogenen Anpassungsfaktoren gebildet wurde. Im Mikrozensus ab 2005 wird ein modifiziertes Verfahren eingesetzt (siehe Afentakis und Bihler 2005; Statistisches Bundesamt 2008).

Die Hochrechnungsfaktoren des Mikrozensus (Variablen ef570ff.) stehen im Scientific Use File und im Campus File für Auswertungen auf der Personenebene (ef750), für Haushalte und Familien (ef751) und Wohnungen (ef761) sowie für die Substichprobe (Personen: ef755; Haushalte und Familien: ef756) zur Verfügung. Aufgrund der Substichprobenziehung werden mit diesen Variablen gewichtete Auwertungen nicht exakt mit Verteilungen der Populationsdaten übereinstimmen. Das Campus File enthält darüber hinaus Hochrechnungsfaktoren (ef750g, ef751g, ef761g), die nachträglich noch nach Bundesland, Geschlecht und Staatsangehörigkeit an Verteilungen des Original-Mikrozensus angepasst wurden.[13] Bei Gewichtung mit diesen Hochrechnungsfaktoren sollten sich die mit dem Campus File erzielten Ergebnisse nur unwesentlich von den veröffentlichten Gesamtwerten der amtlichen Statistik unterscheiden.[14] Tabelle 4 beschreibt für die Bevölkerung am Hauptwohnsitz (ef505<3) das arithmetische Mittel sowie das erste und neunte Dezil der Gewichte pro Anpassungsklasse im Campus File.[15]

Tabelle 4: Statistische Kennziffern zum Hochrechnungsfaktor ef750g[+)] im Campus File Mikrozensus 2002

Anpassungsklassen	Mittelwert	1. Dezil	9. Dezil	n
Deutsche Männer	1,15	1,08	1,23	11.051
Deutsche Frauen	1,12	1,05	1,20	12.159
Ausländische Männer	1,68	1,29	2,12	777
Ausländische Frauen	1,62	1,30	2,08	720
Zeit-/Berufssoldaten	1,45	1,01	2,00	49
Wehrpflichtige	1,44	0,93	1,83	22
Ingesamt	1,16	1,06	1,25	24.778

Quelle: Campus File Mikrozensus 2002 (Auswahl: Bevölkerung am Hauptwohnsitz; [+)] ef750g*0,035)

Stata stellt für die Anwendung der gebundenen Hochrechnung im Kommando `svyset` die Option `poststratification` bereit, deren Einsatzmöglichkeiten im Folgenden beschrieben werden. Die Verwendung von Gewichten, die aus der Anpassung an die Bevölkerungsfortschreibung resultieren, kann alternativ als Regressionsschätzung interpretiert werden. Dieses Verfahren kann auch für die Mikrozensusdaten ab 2005 eingesetzt werden, in denen keine

[13] Die Hochrechnungsfaktoren für Haushalte und Familien (ef751) sind als arithmetischer Mittelwert der Hochrechnungsfaktoren für Personen (ef750) im Haushalt für alle Personen eines Haushalts gleich. Für die Variable ef751g im Campus File trifft diese Eigenschaft allerdings nicht zu.

[14] Allerdings basieren die Fehlerrechnungen des Statistischen Bundesamtes nur auf den Stichprobenwerten. Die Poststratifikation bleibt hierbei außer Acht.

[15] Da die Variable ef750g auf 100 Personen in der Population hochrechnet, wurde ef750g mit dem Auswahlsatz von 3,5 % multipliziert.

disjunkten Anpassungsklassen mehr vorliegen, da die Anpassung ab 2005 auf unterschiedlichen regionalen Ebenen an getrennte Randverteilungen der Populationsdaten bzw. sogenannte Hilfsmerkmale mittels Regressionsschätzung (Kalibrierung) erfolgt (siehe Afentakis und Bihler 2005). Die Regressionsschätzung wird im nächsten Abschnitt dargestellt.

Mit der Anweisung `poststrata(varname)` wird die Variablenkombination Anpassungsschicht (Bundesland * (Geschlecht (m / w), Staatsangehörigkeit (d / ¬d), Bevölkerungsgruppe (Zivilbevölkerung / Soldaten / Wehrpflichtige)) benannt, nach der die Anpassung durchgeführt wurde und für die Informationen im Campus File verfügbar sind.[16] Die Gewichtungsvariable zur gebundenen Hochrechnung kann in Stata nicht direkt eingesetzt werden, sondern mit der Option `postweight(varname)` sind die den Anpassungsgruppen zugehörigen Populationswerte (`M_k`) zu übergeben, die zuvor mit der Gewichtungsvariablen berechnet werden müssen. Als Gewichtungsvariable wird im Folgenden der Personenfaktor ef750g des Campus File herangezogen, da davon auszugehen ist, dass infolge der nachträglichen Anpassung an Verteilungen des Original-Mikrozensus evtl. Abweichungen zu Verteilungen der Population minimal sind. Das interessierende Merkmal und die Subpopulation sind wie im Beispiel 4 abgegrenzt.

Stata (2007a: 51-52) schätzt bei Poststratifikation den Gesamtwert und die Varianz unter Verwendung des Gewichtes *w**, welches das Verhältnis des geschätzten Wertes für den Umfang in der Anpassungsschicht *k* in der Population M_k („Soll“) zum designgewichteten Umfang der Gruppe in der Stichprobe $\hat{M}_k$ („Ist“) für jede Person *j* wiedergibt (siehe Variable `w_post` im Beispiel 5 unten):

$$w_j^* = \sum_{k=1}^{L_p} I_{P_k}(j) \frac{M_k}{\hat{M}_k} w_j \qquad \text{mit} \qquad \hat{M}_k = \sum_{j=1}^{m} I_{P_k}(j) w_j$$

wobei L_p für die Anzahl der Anpassungsschichten steht und $I_{P_k}(j)$ eine Indikatorfunktion für die Zugehörigkeit einer Person zu einer Anpassungsschicht ist.

Bei der Schätzung wird das Gewicht w_j^* eingesetzt, womit die in einer Anpassungsschicht hinsichtlich der verwendeten demografischen Variablen ermittelte allgemeine Über- oder Untererfassung („Soll durch Ist“) auf die interessierende Variable übertragen wird. Dabei ist

[16] Die regionale Anpassungsschicht ist aus Datenschutzgründen nicht identifizierbar. Zwar könnte wie bei der Schichtung auch hier zusätzlich die Gemeindegrößenklasse als Proxy-Information verwendet werden, da dies jedoch mit gering besetzten Anpassungsschichten verbunden sein dürfte, wird darauf verzichtet.

darauf zu achten, dass sich die Hochrechnungsfaktoren für die interessierende Variable und die Subpopulation bzw. Population unterscheiden können (s. u.).

Der Gesamtwert bei Poststratifikation wird geschätzt mit

$$\hat{Y}^P = \sum_{j=1}^{m} w_j^* y_j$$

Würde man für den gewichteten Gesamtwert die oben bei der Designgewichtung verwendete Varianzformel und das entsprechende Stata-Programm einsetzen, wäre dies aufgrund der zusätzlichen Variation der Gewichtungsvariablen mit einer höheren Varianz als bei der Designgewichtung verbunden. Außerdem wird die Idee, dass mit der Anpassung (geschätzte) Populationswerte vorliegen, die für eine Varianzschätzung nutzbar sind, nicht umgesetzt. Greift man diese auf, ist allerdings zu berücksichtigen, dass die Umfänge der Anpassungsschichten in den anonymisierten Daten wie das interessierende Merkmal zufällig sind, bzw. einen Stichprobenfehler aufweisen. Die Varianzfunktion wird deshalb komplex, kann aber durch eine Näherung vereinfacht werden. Für die Varianzschätzung wird eine Hilfsvariable (score variable) eingesetzt.

Als asymptotische Varianz wird in Stata (2007a: 155f.) die Varianz dieser Hilfsvariablen z_j verwendet. Je näher die Stichprobenwerte y_j bei den Populationswerten einer Anpassungsschicht $\hat{Y}_k / \hat{M}_k$ liegen, umso kleiner wird der Standardfehler. Führt man die Varianzschätzung für Gesamtwerte mit dieser Hilfsvariablen durch, erhält man das gleiche Ergebnis wie mit der Varianzschätzung bei Poststratifikation (s. u.).

$$z_j\left(\hat{Y}^P\right) = \sum_{k=1}^{L_p} I_{P_k}(j) \frac{M_k}{\hat{M}_k}\left(y_j - \frac{\hat{Y}_k}{\hat{M}_k}\right) \qquad \text{mit} \qquad \hat{Y}_k = \sum_{j=1}^{m} I_{P_k}(j) w_j y_j$$

Beispiel 5: Gesamtwerte mit gebundener Hochrechnung (Poststratifikation)

```
* Variablenabgenzungen analog zu Beispiel 1b und 3a
* Gesamtwert designgewichtet
svyset ef3 [pw = w], strata(schicht) fpc(f) ///
                     single(certainty)
svy linearized, subpop(sub) : total y , noheader
```

```
        |             Linearized
        |      Total   Std. Err.     [95% Conf. Interval]
  -----+-------------------------------------------------
     y |    2874286   94763,38       2688532     3060040
  -------------------------------------------------------
  Note: strata with single sampling unit treated as certainty units.
```

```
matrix t_d = e(b)
svmat  t_d
matrix v_d = e(V)
```

```
svmat  v_d
disp "CV (%) = " (v_d^.5 * 100)/t_d
  CV (%) = 3,2969365

* Vorbereitung Poststratifikation
* 1a) Anpassung: Geschl. * Staatsangeh., Sold., Wehrpfl.
gen anp=(1*(ef32==1 & ef52==1 & ef127~=9 & ef127~=10)) ///
      + (2*(ef32==2 & ef52==1)) ///
      + (3*(ef32==1 & ef52~=1)) ///
      + (4*(ef32==2 & ef52~=1)) ///
      + (5*(ef32==1 & ef52==1 & ef127==9)) ///
      + (6*(ef32==1 & ef52==1 & ef127==10))
lab var anp "Anpassungsklassen"
lab def anp 1 "Deutsche Maenner"  2 "Deutsche Frauen" ///
            3 "Ausl. Maenner"     4 "Ausl. Frauen" ///
            5 "Zeit-/Berufssold." 6 "Wehrpflichtige"
label val anp anp

* 1b) Proxy Anpassungsschicht
gen anpschicht=sub*(ef1*10+anp) /* nur für Subpop */
lab var anpschicht "Anpassungschicht - Proxy"

* 2)  Populationsdaten pro Anpassungsschicht
gen w_anp = ef750g*100*(ef505<=2) // Bevölkerung am Hauptwohnsitz
egen    M_k = total(w_anp), by(anpschicht)

* Total gebundene Hochrechnung
svyset ef3 [pw = w], strata(schicht) fpc(f) single(certainty) ///
                     poststrata(anpschicht) postweight(M_k)
svy linearized, subpop(sub) : total y , noheader

  Survey: Total estimation
  Number of strata =      79        Number of obs    =     25137
  Number of PSUs   =   10707        Population size  =  82391641
  N. of poststrata =      87        Subpop. no. obs  =     21193
                                    Subpop. size     =  70405706
                                    Design df        =     10628
       |             Linearized
       |      Total   Std. Err.     [95% Conf. Interval]
  -----+-------------------------------------------------
     y |    3428499   111236,1       3210455     3646542
  -------------------------------------------------------
  Note: strata with single sampling unit treated as certainty units.

matrix t_p = e(b)
svmat  t_p
matrix v_p = e(V)
svmat  v_p
disp "CV (%) = " (v_p^.5 * 100)/t_p
  CV (%) = 3,244454

* 3) Umsetzung der Stata-Formeln
egen M_Dach_k = total(w*sub), by(anpschicht)
gen w_post = M_k/M_Dach_k*w                  /* w*_j  */
replace w_post = 0 if sub==0
gen y_post = w_post*y

* Gesamtwert y_post: poststratifiziert
* <=> Abweichung zur Gewichtung mit ef750g, z.B. mit:
      tab y if sub [iw=w_anp]
```

```
table y if y & sub, c(sum y_post sum w_anp)

   ------------------------------------
           y | sum(y_post)   sum(w_anp)
   ----------+-------------------------
           1 |     3428499      3431212
   ------------------------------------

* V(Y_post) für Hilfsmerkmal (score variable) z
egen Y_Dach_k = total(w*y*sub), by(anpschicht)
gen z = sub * (M_k/M_Dach_k) * (y - (Y_Dach_k/M_Dach_k))
replace z=0 if z==.
svyset ef3 [pw = w], strata(schicht) fpc(f) ///
                     single(certainty)
svy linearized, subpop(sub) : total z, noheader
{...}
matrix v_z = e(V)
svmat  v_z
disp "s.e. (z) = " v_z^.5
   s.e. (z) = 111236,06
```

Wie das obige Beispiel zeigt, schließt das 95%-Konfidenzintervall (2.688.532, 3.060.040) des designgewichteten Gesamtwertes von 2.874 Tsd. Erwerbslosen nicht das veröffentlichte Ergebnis von 3.486 Tsd. Erwerbslosen (Statistisches Bundesamt 2003b: 161) ein. Dagegen wird durch die gebundene Hochrechnung mit 3.428,5 Tsd. Erwerbslosen ein Gesamtwert geschätzt, der nahe beim Referenzwert liegt und dessen Konfidenzintervall diesen Referenzwert enthält. Der relative Standardfehler liegt mit 3,2 Prozent nur geringfügig unter dem relativen Standardfehler der designgewichteten Schätzung von 3,3 Prozent. Dies war kaum anders zu erwarten, da die demografischen Anpassungsmerkmale nur schwach mit dem Merkmal Erwerbslos korreliert sind.

Die Umsetzung der Stata-Formeln und die Berechnung „von Hand“ am Ende des Beispielprogramms dokumentiert, wie die Hilfsvariable z_j für die Varianzschätzung eingesetzt werden kann.

Allerdings fällt auf, dass der Gesamtwert der gebundenen Hochrechnung von der einfachen gewichteten Auswertung abweicht (siehe in obiger Tabelle „sum(y_post)“ vs. „sum(w_anp)“), da mit der Gewichtungsvariablen w_j^* (*w_anp*) für M_k der Populationsumfang in der Anpassungsschicht *k* für die Bevölkerung am Hauptwohnsitz berechnet wurde, jedoch die Hochrechnungsfaktoren für die interessierende Variable davon verschieden sind. Um dieses Problem zu umgehen, können die Gewichte für die interessierende Variable verwendet werden.

```
/* Problem: Gesamtwert der gebundenen Hochrechnung stimmt nicht mit
   gewichteter Tabellierung überein. Lösung: Verwendung des
   Korrekturfaktors für y-Werte und Übertragung auf (Sub-)Population. */
egen wy = total(ef750g*0.035*y), by(anpschicht) // "SOLL"
```

```
egen ny = total(y), by(anpschicht)              // "IST"
gen g_y = wy/ny // Korrekturfaktor der y-Werte
recode g_y (.=0)
gen w_y = g_y*w  // Endgewicht = Korrekturfaktor * Designgewicht
egen M_k2 = total(w_y), by(anpschicht)

svyset ef3 [pw=w], strata(schicht) fpc(f) single(certainty) ///
                   poststrata(anpschicht) postweight(M_k2)
svy linearized, subpop(sub) : total y

  Survey: Total estimation
  Number of strata =      79        Number of obs    =     25137
  Number of PSUs   =   10707        Population size  =  69718181
  N. of poststrata =      87        Subpop. no. obs  =     21193
                                    Subpop. size     =  69718181
                                    Design df        =     10628

             |             Linearized
             |      Total   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
           y |    3431212   111350,4       3212945     3649480
-------------------------------------------------------------
  Note: strata with single sampling unit treated as certainty units.

matrix t_k = e(b)
svmat  t_k
matrix v_k = e(V)
svmat  v_k
disp "CV (%) = " (v_k^.5 * 100)/t_k
  CV (%) = 3,2452192
```

Mit der Übertragung dieser Korrekturfaktoren für das interessierende Merkmal ($y = 1$) auf die gesamte Population und Subpopulation (inkl. $y = 0$) ändern sich auch die Schätzungen der entsprechenden Größen (vgl. die jeweiligen Angaben zu „Population size“ und „Subpop. size“). Die so ermittelte Standardabweichung unterscheidet sich von den obigen Ergebnissen nur geringfügig. Dagegen ist der Variationskoeffizient beider Schätzungen gleich. Insgesamt betrachtet, erscheint das zweite Vorgehen angemessen, weil die veröffentlichten Ergebnisse repliziert werden.

Alternativ könnte man zu dem für plausibler gehaltenen Gesamtwert von $\hat{Y}^P$ = 3.431.212 die formal „korrekte“ erste Varianzschätzung $\hat{V}\left(\hat{Y}^P\right)$ = 111.236,10 übernehmen. Mittels Berechnung des Variationskoeffizienten zur ersten Schätzung (cv = 111.236,10 / 3.428.499 = 0,0324) wird der Standardfehler auf $\hat{V}'\left(\hat{Y}^P\right)$ = 0,0324 * 3.431.212 = 111.324,12 geschätzt.

4.5 Gebundene Hochrechnung mittels Regressionsschätzung

Die Verwendung von Gewichten, die im Wesentlichen aus der Anpassung der Mikrozensus-Fallzahlen für bestimmte Merkmalskombinationen an die Bevölkerungsfortschreibung resultieren, kann als Regressionsschätzung interpretiert werden. Im Kapitel zur Poststratifikation

wurden die in Form des Hochrechnungsfaktors vorliegenden Populationsdaten verwendet, um eine bessere oder zumindest eine mit der laufenden Bevölkerungsfortschreibung konsistentere Schätzung zu erreichen. Mit der Regressionsschätzung wird dieses Vorgehen verallgemeinert, insofern angenommen wird, dass die Populationsdaten bzw. Hilfsvariablen (**x**) einen statistischen Einfluss auf die interessierende Variable (*y*) haben, der durch eine Regression modelliert werden kann (Särndal et al. 1997: 245ff.).

In der Stichprobe *s* liegen somit für die Personen *k* Hilfsvariablen $\left(x_{k1},...,x_{kp}\right)$ vor, deren Gesamtwerte mit Designgewichtung (Variable d_k; Horvitz-Thompson-Schätzung) geschätzt werden: $\hat{t}_{x,HT} = \sum_{k\in s} d_k \mathbf{x}_k$. In der Population *U* sind die Gesamtwerte bekannt: $t_x = \sum_{k\in U} \mathbf{x}_k$.

Hätte man die Daten für die Grundgesamtheit und betrachtet man die Werte des interessierenden Merkmals *y* als Zufallsvariable, könnte man den postulierten Zusammenhang durch folgendes Modell ξ darstellen:

$$E_\xi\left(y_k\right) = \beta' x_k \qquad V_\xi\left(y_k\right) = \sigma_k^2$$

Das Ziel besteht darin, unter Verwendung der Populationsdaten zu den Hilfsvariablen $\hat{t}_{x,HT}$ den Gesamtwert der interessierenden Variablen *y* unter Berücksichtigung des Designgewichtes $d_k = 1/\pi_k\hat{\theta}_k$ zu schätzen, wobei π_k die Inklusionswahrscheinlichkeit und $\hat{\theta}_k$ die für die Korrektur der Haushaltsausfälle geschätzte Antwortwahrscheinlichkeit darstellen.[17] Die neu zu bestimmenden Gewichte $w_k = g_k * d_k$ sind dabei so zu wählen, dass die mit w_k gewichteten Hilfsvariablen die Populationsverteilungen wiedergeben: $t_x = \sum_{k\in S} w_k \mathbf{x}_k$, und die Gewichte w_k möglichst nahe bei den Designgewichten d_k liegen.

Der verallgemeinerte Regressionsschätzer eines Gesamtwertes (Total) ist in Matrixnotation:

$$\begin{aligned}\hat{t}_{y,reg} &= \sum_{k=1}^{n}\left(1+\left(\mathbf{t}_x-\hat{\mathbf{t}}_{x,HT}\right)'\left(\sum d_k\mathbf{x}_k\mathbf{x}_k'\right)^{-1}\mathbf{x}_k\right)d_k y_k \\ &= \hat{t}_{y,HT}+\hat{\mathbf{B}}'\left(\mathbf{t}_x-\hat{\mathbf{t}}_{x,HT}\right)\end{aligned}$$

mit den geschätzten Regressionskoeffizienten

$$\hat{\mathbf{B}} = \left(\sum_{k=1}^{n} d_k\mathbf{x}_k\mathbf{x}_k'\right)^{-1}\left(\sum_{k=1}^{n} d_k\mathbf{x}_k y_k\right)$$

[17] In den anonymisierten Daten liegt der Kompensationsfaktor für Haushaltsausfälle ($1/\pi_k$) nicht vor, sondern ist im endgültigen Hochrechnungsfaktor der gebunden Hochrechnung (ef750 usw.) enthalten.

Werden die Hilfsmerkmale $\hat{\mathbf{t}}_{x,HT}$ bei designgewichteter Hochrechnung der Stichprobe gegenüber dem Populationswert $\mathbf{t}_x$ unterschätzt und sind sie mit der interessierenden Variablen y positiv korreliert, dann wird auch der zu schätzende Gesamtwert $\hat{\mathbf{t}}_{y,HT}$ unterschätzt und durch die Regressionsschätzung entsprechend korrigiert. Das Ausmaß der Korrektur hängt von zwei Faktoren ab. Erstens vom Regressionskoeffizienten $\hat{\mathbf{B}}$, der mittels linearer Regression der zur Anpassung verwendeten Hilfsmerkmale **x** auf die interessierende Variable y geschätzt wird. Je enger die Hilfsmerkmale mit der interessierenden Variable korreliert sind, umso stärker wird die Korrektur ausfallen. Zweitens hängt die Korrektur davon ab, wie nahe die mit der Stichprobe geschätzten Gesamtwerte der Hilfsmerkmale $\hat{\mathbf{t}}_{x,HT}$ bei den bekannten Gesamtwerten der Population $\mathbf{t}_x$ liegen.

Der im Gewicht w_k enthaltene Korrekturfaktor g_k, der in die Schätzung der interessierenden Variable einfließt, gibt den Beitrag der Hilfsmerkmale zur Reduktion von Abweichungen zwischen Stichprobe und Population wieder und kann, wenn keine weiteren Restriktionen vorliegen,[18] direkt berechnet werden:

$$g_k = 1 + \left(\mathbf{t}_x - \hat{\mathbf{t}}_x\right)' \left(\sum \frac{\mathbf{x}_k \mathbf{x}_k'}{\pi_k \hat{\theta}_k} \right)^{-1} \mathbf{x}_k$$

Die Korrekturfaktoren g_k für disjunkte Anpassungsschichten liegen für die anonymisierten Mikrozensusdaten bis 2004 bereits in Form der Gewichtungsvariablen zur gebundenen Hochrechnung vor und können für die Berechnung der Populationsdaten herangezogen werden. Es kann das Group Mean Model verwendet werden (Särndal et al. 1997: 264ff.); vgl. zum Mikrozensus die ausführliche Darstellung in Rendtel und Schimpl-Neimanns (2001: 103f.).

Unter der Annahme, dass die Anpassungsgruppen bzw. –schichten hinsichtlich der Unter- und Übererfassungen homogen sind und sich dies in gruppenspezifischen „Soll durch Ist"-Faktoren ausdrückt, lässt sich die Grundgesamtheit U in disjunkte Teilmengen $U_g, g \in \{1,...,G\}$ zerlegen. Für die Elemente innerhalb jeder Gruppe gilt:

$$E_\xi\left(y_k\right) = \mathbf{x}_g' \boldsymbol{\beta} \qquad V_\xi\left(y_k\right) = \sigma_g^2 \qquad k \in U_g$$

Der Regressionskoeffizient β_g wird auf Basis der Stichprobe mit einer linearen Regression geschätzt, kann aber aufgrund disjunkter Anpassungsschichten auch einfacher als arithmeti-

[18] Im Mikrozensus ab 2005 werden für die Korrekturfaktoren Unter- und Obergrenzen bestimmt, sodass die Berechnung der Korrekturfaktoren iterativ durchgeführt werden muss (Afentakis und Bihler 2005).

sches Mittel der interessierenden Variablen in den einzelnen Anpassungsschichten berechnet werden (siehe Variable *B_Dach* im Beispiel 6):

$$\hat{B}_g = \frac{1}{\hat{N}_g} \sum_{k \in s_g} \hat{y}_k = \overline{y}_{s_g}$$

Der Regressionsschätzer des Gesamtwertes lässt sich in diesem Fall als „gewichtetes Mittel" darstellen:

$$\hat{t}_{y,reg} = \sum_{k=1}^{n} w_k y_k$$

Analog zum Vorgehen bei Poststratifikation im vorigen Kapitel wird die Varianzschätzung auch hier mit einer Hilfsvariablen durchgeführt. Als asymptotische Varianz wird die Varianz der Hilfsvariablen u_k verwendet (siehe Variablen *u1-u3* im Beispiel 6).

$$u_k = \frac{N_g}{\hat{N}_g}\left(y_k - \hat{B}_g\right) = w_k\left(y_k - \overline{y}_{s_g}\right) \qquad k \in s_g$$

Je besser also das interessierende Merkmal y von den bei der Anpassung berücksichtigten Hilfsvariablen statistisch erklärt wird, d. h. je kleiner die Residuen der Regression von den y-Werten auf die Hilfsvariablen bzw. je kleiner die Differenz ($y_k - \overline{y}_{sg}$), umso geringer ist die Varianz des Regressionsschätzers im Vergleich zur designgewichteten Schätzung.

Die Varianzschätzung wird designbasiert durchgeführt. Da das statistische Modell ξ nur für die Herleitung des Populationsschätzers dient, wird der Regressionsschätzer als modellgestützt („model assisted") bezeichnet.

Selbst wenn die Modellannahmen kritisch zu betrachten sind, halten Särndal et al. (1997: 239) fest: „If the population data are well described by the assumed model, the regression estimator normally will bring about a large variance reduction, as compared to the π estimator [HT-Schätzer; B.S-N.]. If the population is not well described by the model, the improvement on the π estimator may be modest, but the regression estimator still guarantees approximate unbiasedness."

Gleichwohl ist zu beachten, dass sowohl die Populationsdaten als auch die Stichprobendaten nicht frei von systematischen Fehlern sind. Einerseits ist zur Qualität der Bevölkerungsfortschreibung bekannt, dass sie zu hohe Ausländerzahlen aufweist (siehe Opfermann et al., 2006), sodass mit der Anpassung des Mikrozensus an Ergebnisse der Bevölkerungsfortschrei-

bung auch deren Fehler auf den Mikrozensus übertragen werden können.[19] Andererseits ist zum Erwerbsstatus im Mikrozensus bekannt, dass die Zahlen der Erwerbstätigen und Erwerbslosen im Vergleich zu anderen Quellen untererfasst sind und diese Untererfassung durch die Hochrechnungsfaktoren nicht ausgeglichen wird (vgl. Statistisches Bundesamt 2006).

Tabelle 5 enthält die Ergebnisse der Regressionsschätzungen des Beispiels 6, wenn [A] der Hochrechnungsfaktor ef750g, wie zuvor im Beispiel 5, als auf die Population insgesamt bezogener Korrekturfaktor (*g1*) oder [B] auf die interessierende Variable bezogen (*g2*) sowie [C] (nach Normierung auf die Stichprobe) direkt (*g3*) verwendet wird.

Tabelle 5: Ergebnisse der Regressionsschätzungen mit verschiedenen Gewichtungsvarianten (Beispiel 6)

	Korrekturfaktor *g*	Gesamtwert	Standardabw.	CV (%)
A	*g1*: Korrekturfaktoren der (Sub-) Population pro Anpassungsschicht	3.428.499	111.236	3,24
B	*g2*: Korrekturfaktoren der y-Werte pro Anpassungsschicht	3.431.212	111.350	3,25
C	*g3*: Korrekturfaktoren der y-Werte direkt	3.431.212	111.379	3,25

Die den Versionen A und B entsprechenden Verfahren führen hinsichtlich der geschätzten Gesamtwerte und Standardfehler zu den gleichen Ergebnissen wie in Beispiel 5 (siehe S. 34 und S. 36).[20] Version [C] resultiert wie [B] im gleichen gewichteten Gesamtwert. Die Version C hat nicht nur den „kosmetischen" Vorteil, dass die Schätzung mit einer einfachen gewichteten Auswertung übereinstimmt. Sie führt infolge der Variation des g-Gewichtes (bzw. ef750g) innerhalb der Anpassungsschichten zu einem etwas größeren Standardfehler. Da aber die relativen Standardfehler bzw. Variationskoeffizienten gleich sind, können die Unterschiede insgesamt als vernachlässigbar betrachtet werden.

In den meisten Fällen dürfte mit der Option „poststratification" die einfachste Varianzschätzung bei gebundener Hochrechnung des Mikrozensus bis einschließlich 2004 erreichbar sein. Bei den Mikrozensen ab 2005, in denen keine disjunkten Anpassungsschichten mehr verwendet werden, kann die eingangs beschriebene verallgemeinerte Regressionsschätzung benutzt werden.

[19] Allerdings dürften diese bei Weitem nicht in dem Umfang von rund 60 Prozent auftreten, den die Korrekturfaktoren für Ausländer in Tabelle 4 widerspiegeln. Z. B. reduzierte sich die Zahl der ausländischen Bevölkerung im Mikrozensus 1987 nach der revidierten Hochrechnung auf Basis der Volkszählung 1987 lediglich um zwölf Prozent (Heidenreich 1989).

[20] Wendet man die oben in Matrixnotation dargestellte Regressionsschätzung (siehe Formeln zu $\hat{\mathbf{B}}$ und g_k) an und gewichtet mit dem Korrekturfaktor *g3*, erhält man die gleichen Ergebnisse wie mit Version A.

Beispiel 6: Gesamtwerte mit gebundener Hochrechnung (Regressionsschätzung)

```
* Variablenabgenzungen analog zu Beispiel 1 und 5
gen anpschicht=sub*(ef1*10+anp)

* Berechnen des Regressionskoeffizienten
* 1c (1) OLS-Regression (für Subpopulation) mit Regression y "Erwerbslos"
*        auf X "Hilfsmerkmale" bzw. "Anpassungsschicht"
*        xi: regress y i.anpschicht if sub,
*        predict B_Dach if e(sample), xb
*        replace B_Dach = 0 if B_Dach==.
* Einfacher, da disjunkte Anpassungsschichten:
egen B_Dach = mean(y), by(anpschicht)

* [A] Gewichtung analog Beispiel 5 (w_anp)
*  3) Berechnung d. gewichteten Gesamtwertes über Anp.schichten
*       ef750g rechnet auf 100 in Population hoch
egen M_k   = total(ef750g*100*(ef505<=2)), by(anpschicht)

* Populationswerte geb. Hochrechnung
egen M_Dach_k = total(d*(ef505<=2)), by(anpschicht)

* Populationswerte designgewichtet
gen g1 = M_k/M_Dach_k     // Korrekturfaktor g_k
replace g1 = 0 if sub==0
gen  w1 = d*g1            // Endgewicht w_k
egen t_y1 = total(y*w1)

* Varianzschätzung mit Hilfsvariable u
gen u1 = g1*(y-B_Dach)

* [B] Gewichtung analog Beispiel 5 (w_y)
egen wy = total(ef750g*0.035*y), by(anpschicht)
egen ny = total(y), by(anpschicht)

gen  g2 = wy/ny // Korrekturfaktor der y-Werte
recode g2 (.=0)

gen w2 = g2*d  // Endgewicht = Korrekturfaktor * Designgewicht
egen t_y2 = total(y*w2)

gen u2    = g2*(y-B_Dach)

* [C] Gewichtung mit ef750 direkt (ohne Normierung auf 1%-MZ); siehe
*     frühere Programme (Rendtel/Schimpl-Neimanns 2001); z.B. varmz_a.do
gen  g3 = ef750g*0.035*sub
gen  w3 = d*g3
egen t_y3 = total(y*w3)
gen u3    = g3*(y-B_Dach)

disp "Gewichteter Gesamtwert t_y1 = " t_y1
  Gewichteter Gesamtwert t_y1 = 3428498,8
disp "Gewichteter Gesamtwert t_y2 = " t_y2
  Gewichteter Gesamtwert t_y2 = 3431212
disp "Gewichteter Gesamtwert t_y3 = " t_y3
  Gewichteter Gesamtwert t_y3 = 3431212,3

svyset ef3 [pw = d], strata(schicht) fpc(f) single(certainty)
* "total" jeweils irrelvant
svy linearized, subpop(sub) : total u1 u2 u3 , noheader
```

	Linearized			
	Total	Std. Err.	[95% Conf.	Interval]
u1	,0159868	111236,1	-218043,5	218043,5
u2	,0187489	111350,4	-218267,5	218267,6
u3	2713,289	111378,9	-215610,2	221036,8

5 Verhältniswerte

Wie bereits zu den Gruppenvergleichen im Abschnitt 4.3 und in Beispiel 4 angesprochen, spielen in der Forschungspraxis Verhältnis- oder Anteilswerte häufig eine größere Rolle als die oben betrachteten Gesamtwerte. Da bei dieser Verhältnisschätzung sowohl im Zähler- als auch im Nennermerkmal Stichprobenfehler auftreten, muss deren Kovarianz berücksichtigt werden, sodass die Berechnung nicht ganz so einfach ist wie bei Gesamtwerten.

5.1 Designbasierte Schätzung

Für den Verhältnisschätzer *R* der Gesamtwerte von zwei Merkmalen $\hat{t}_y$ und $\hat{t}_x$ (bzw. $\hat{Y}$ und $\hat{X}$)

$$\hat{R} = \hat{t}_y / \hat{t}_x$$

wird bei Designgewichtung zur Berechnung der asymptotischen Varianz

$$\hat{V}\left(\hat{R}\right) = \frac{1}{\hat{X}^2}\left\{\hat{V}\left(\hat{Y}\right) - 2\hat{R}\hat{V}\left(\hat{Y},\hat{X}\right) + \hat{R}^2\hat{V}\left(\hat{X}\right)\right\}$$

wieder eine Hilfsvariable (score variable) benutzt (Stata 2007a: 155):

$$z_j\left(\hat{R}\right) = \frac{y_j - \hat{R}x_j}{\hat{X}}$$

Damit kann die im Unterabschnitt 4.4.1 (Seite 16) genannte Formel zur Varianzschätzung von Gesamtwerten mit dieser Hilfsvariablen auch für die Varianzschätzung von Verhältniswerten verwendet werden.

Die Verhältnisschätzung soll am Beispiel der bereits im Unterabschnitt 4.3 behandelten Erwerbslosenquote, also dem Anteil der Erwerbslosen an den Erwerbspersonen (Erwerbstätige und Erwerbslose) gezeigt werden, die nach Region (West- vs. Ostdeutschland) und drei Altersgruppen (15-24, 25-54 und 55-65 Jahre) differenziert werden. Neben der freien Hochrechnung bzw. designbasierten Schätzung wird auch eine gebundene Hochrechnung (Poststratifikation) vorgenommen.

Die Erwerbslosenquote ist, wie in Beispiel 4 bereits festgestellt, in Ostdeutschland rund dreimal höher als in Westdeutschland (18,2 % vs. 6,1 %; siehe Seite 29). Im folgenden Beispiel 7 ist anhand der ersten Auswertung nach Altersgruppen zu sehen, dass Jugendliche und ältere Personen mit rund elf Prozent eine höhere Erwerbslosenquote aufweisen als die mittlere Altersgruppe (7,5 %). Mit der zweiten Auswertung werden die alters- und regionenspezifi-

schen Risiken erkennbar. Während beispielsweise von den jüngeren westdeutschen Erwerbspersonen rund neun Prozent erwerbslos sind, trifft dies bei den ostdeutschen Jugendlichen auf 22 Prozent zu. Die unter Berücksichtigung des Stichprobendesigns durchgeführten Wald-Tests auf Gleichheit der Anteile ergeben statistisch signifikante Unterschiede.[21]

Beispiel 7: Erwerbslosenquote mit Designgewichtung

```
* Variablenabgenzungen analog zu Beispiel 4
* sub: Bevölkerung am Hauptwohnsitz, 15-65 Jahre, Erwerbspersonen
gen sub = ef505>=1 & ef505<=2 & ///
          ef30>=15 & ef30<=65 & ef504>=1 & ef504<=2
gen y = (ef504==2) * sub // Erwerbslose
gen x = (ef504>=1 & ef504<=2) * sub // Erwerbspersonen

* Erwerbslosenquote nach Altersgruppen
svyset ef3 [pw=w] , strata(schicht) fpc(f) singleunit(certainty)
svy linearized, subpop(sub) : ratio (Erwerbslosenquote: y/x), over(alter)
estat effects, deft

   Erwerbslos~e: y/x
      _subpop_1: alter = 15-24
      _subpop_2: alter = 25-54
      _subpop_3: alter = 55-65
  ------------------------------------------------
               |             Linearized
          Over |      Ratio   Std. Err.       DEFT
  -------------+----------------------------------
  Erwerbslos~e |
     _subpop_1 |   ,1166078   ,0087993     1,03129
     _subpop_2 |   ,0752497   ,0029034     1,04471
     _subpop_3 |   ,1114149   ,0085067     1,03407
  ------------------------------------------------

* Erwerbslosenquote nach Region und Altersgruppen
svy linearized, subpop(sub): ratio (Erwerbslosenquote: y/x),
    over(westost alter)
estat effects, deft

   Erwerbslos~e: y/x
           Over: westost alter
      _subpop_1: West 15-24
      _subpop_2: West 25-54
      _subpop_3: West 55-65
      _subpop_4: Ost 15-24
      _subpop_5: Ost 25-54
      _subpop_6: Ost 55-65
```

[21] Im Fall multiplen Testens stellt Stata mit dem Kommando `test` und der Option `mtest` u. a. eine Bonferroni-Korrektur bereit.

```
------------------------------------------------
             |             Linearized
        Over |      Ratio   Std. Err.       DEFT
-------------+----------------------------------
Erwerbslos~e |
   _subpop_1 |   ,0883929   ,0086881     1,02427
   _subpop_2 |    ,054173   ,0027738     1,04502
   _subpop_3 |   ,0791246   ,0079384     1,01362
   _subpop_4 |   ,2237288    ,025269     1,04141
   _subpop_5 |   ,1635003   ,0093316     1,05161
   _subpop_6 |   ,2509091    ,027887     1,06668
------------------------------------------------
```

```
* Wald-Tests
* Erwerbslosenquote 15-24-Jähriger: West = Ost ?
test [Erwerbslosenquote]_subpop_1 = [Erwerbslosenquote]_subpop_4

  Adjusted Wald test
  {...}
         F(  1, 10585) =   25,65
              Prob > F =    0,0000

* Erwerbslosenquote 55-65-Jähriger: West = Ost ?
test [Erwerbslosenquote]_subpop_3 = [Erwerbslosenquote]_subpop_6

  Adjusted Wald test
  {...}
         F(  1, 10585) =   35,11
              Prob > F =    0,0000
```

5.2 Poststratifikation

Bei gebundener Hochrechnung werden für die Verhältnisschätzung

$$\hat{R}^P = \frac{\hat{Y}^P}{\hat{X}^P}$$

die bereits im Abschnitt 4.4 genannten Gewichte verwendet (siehe S. 32). Das Zählermerkmal $\hat{Y}^P$ (und analog das Nennermerkmal $\hat{X}^P$) ist wie folgt definiert:

$$\hat{Y}^P = \sum_{k=1}^{L_p} \frac{M_k}{\hat{M}_k} \hat{Y}_k = \sum_{k=1}^{L_p} \frac{M_k}{\hat{M}_k} \sum_{j=1}^{m} I_{P_k}(j) w_j y_j$$

Für die Varianzschätzung benutzt Stata (2007a: 52) die Hilfvariable

$$z_j\left(\hat{R}^P\right) = \frac{\hat{X}^P z_j\left(\hat{Y}^P\right) - \hat{Y}^P z_j\left(\hat{X}^P\right)}{\left(\hat{X}^P\right)^2}$$

mit $$z_j\left(\hat{Y}^P\right) = \sum_{k=1}^{L_p} I_{P_k}(j) \frac{M_k}{\hat{M}_k} \left(y_j - \frac{\hat{Y}_k}{\hat{M}_k} \right)$$

$z_j\left(\hat{X}^P\right)$ ist analog dazu definiert

Ähnlich zu den obigen Erfahrungen ist auch hier zu beachten, dass die mit dem SVY-Kommando erzielten Ergebnisse aufgrund der Berechnung der Populationswerte bzw. der Gewichtung (siehe den Ausdruck $M_k / \hat{M}_k$ in den obigen Formeln) von einer gewichteten Auswertung mit dem Hochrechnungsfaktor (z. B. mit: `ratio y/x if sub [iw = ef750g]`) abweichen können, mit der der Hochrechnungsfaktor der Subpopulation angewendet wird.

Die Ergebnisse bei gebundener Hochrechnung unterscheiden sich nur gering von den obigen Ergebnissen designgewichteter Daten, wobei die geschätzten Anteile um maximal 0,4 Prozent differieren (Verhältniswerte der Subpopulation West, 15-24 Jahre bei Designgewichtung: 8,8 % (Beispiel 7), bei gebundener Hochrechnung: 9,2 % (Beispiel 8)). Diese Unterschiede hängen vor allem damit zusammen, dass der Hochrechnungsfaktor von Ausländern, die überdurchschnittlich erwerbslos sind, größer ist als bei Deutschen (siehe Tab. 3). Somit tragen je nach Zusammensetzung der interessierenden Populationen hinsichtlich der Anpassungsgruppen bzw. Anpassungsschichten die Hochrechnungsfaktoren auch zu einer Umgewichtung der Verhältniswerte bei. Diese Differenzen können jeweils anders ausfallen.

Beispiel 8: Erwerbslosenquote mit gebundener Hochrechnung (Poststratifikation)

```
* Variablenabgenzungen wie im Beispiel 7
*     Anpassungsschichten (siehe Beispiel 5) {...}
gen anpschicht=sub*(ef1*10+anp) /* "0" nur für Subpop */
*    Populationsdaten pro Anpassungsschicht
egen M_k = total(ef750g*100*(ef505<=2)), by(anpschicht)

* Ratio gebundene Hochrechnung
svyset ef3 [pw = w], strata(schicht) fpc(f) single(certainty) ///
                    poststrata(anpschicht) postweight(M_k)
svy linearized, subpop(sub) : ratio (Erwerbslosenquote: y/x), ///
                              over(alter)
Survey: Ratio estimation
{...}
  Erwerbslos~e: y/x
      _subpop_1: alter = 15-24
      _subpop_2: alter = 25-54
      _subpop_3: alter = 55-65
                |             Linearized
           Over |      Ratio   Std. Err.     [95% Conf. Interval]
  -------------+------------------------------------------------
  Erwerbslos~e |
     _subpop_1 |   ,1189993   ,0089161      ,1015221    ,1364765
     _subpop_2 |   ,0770728   ,0029242      ,0713407    ,0828048
     _subpop_3 |   ,1136798   ,0086668      ,0966913    ,1306684
  ------------------------------------------------------------
  Note: 7 strata omitted because they contain no subpopulation members.
```

```
svy linearized, subpop(sub) : ratio (Erwerbslosenquote: y/x), ///
                              over(westost alter)

  Survey: Ratio estimation
  {...}
   Erwerbslos~e: y/x
            Over: westost alter
       _subpop_1: West 15-24
       _subpop_2: West 25-54
       _subpop_3: West 55-65
       _subpop_4: Ost 15-24
       _subpop_5: Ost 25-54
       _subpop_6: Ost 55-65

                |             Linearized
           Over |      Ratio   Std. Err.     [95% Conf. Interval]
  -------------+------------------------------------------------
  Erwerbslos~e |
     _subpop_1 |   ,0923416   ,0090616      ,0745792    ,1101039
     _subpop_2 |   ,0561843   ,0028613      ,0505756     ,061793
     _subpop_3 |   ,0814993   ,0082296      ,0653677     ,097631
     _subpop_4 |   ,2211733    ,024644      ,1728664    ,2694803
     _subpop_5 |   ,1657103    ,009326      ,1474295     ,183991
     _subpop_6 |    ,253315   ,0278233       ,198776    ,3078539
  ------------------------------------------------------------
  Note: 7 strata omitted because they contain no subpopulation members.

* Wald-Tests
* Erwerbslosenquote 15-24-Jähriger: West = Ost ?
test [Erwerbslosenquote]_subpop_1 = [Erwerbslosenquote]_subpop_4

  Adjusted Wald test
  {...}
         F(  1, 10585) =   24,08
              Prob > F =    0,0000

* Erwerbslosenquote 55-65-Jähriger: West = Ost ?
test [Erwerbslosenquote]_subpop_3 = [Erwerbslosenquote]_subpop_6

  Adjusted Wald test
  {...}
         F(  1, 10585) =   35,08
              Prob > F =    0,0000
```

6 Mittelwerte

6.1 Designbasierte Schätzung

Der Populationsmittelwert wird bei Gewichtung mit dem Designgewicht w_j (siehe zur Gewichtung S. 16) von Stata (2007b: 246-247) geschätzt durch

$$\bar{y} = \hat{Y}/\hat{M} \qquad \text{mit} \quad \hat{Y} = \sum_{j=1}^{m} w_j y_j \quad \text{und} \quad \hat{M} = \sum_{j=1}^{m} w_j$$

Die Hilfsvariable für die Varianzschätzung (siehe dazu auch Abschnitt 5.2) ist

$$z_j(\bar{y}) = \frac{y_j - \bar{y}}{\hat{M}} = \frac{\hat{M} y_j - \hat{Y}}{\hat{M}^2}$$

Im obigen Beispiel 7 zur Erwerbslosenquote entsprach das Merkmal im Nenner der Subpopulation, sodass der Verhältniswert von Y = Zahl der Erwerbslosen zu X = Zahl der Erwerbspersonen dem arithmetischen Mittel von Y gleichkommt.

Im folgenden Beispiel wird das arithmetische Mittel der normalerweise geleisteten Wochenarbeitsstunden für Frauen und Männer in West- und Ostdeutschland geschätzt. Einige Werte sind aus Datenschutzgründen vergröbert und werden für die Auswertung auf die Klassenmitte rekodiert. Als Subpopulation wird die erwerbstätige Bevölkerung am Hauptwohnsitz gewählt.

Die Analyse ergibt, dass sich die durchschnittlichen Arbeitszeiten von Männern in West- und Ostdeutschland mit etwas mehr als 40 Stunden nur unwesentlich unterscheiden. Dagegen liegt die Wochenarbeitszeit von Frauen in Ostdeutschland mit rund 35 Stunden deutlich höher als von Frauen in Westdeutschland, die durchschnittlich rund 31 Stunden arbeiten. Die Schätzungen der Mittelwerte bei freier und gebundener Hochrechnung sind praktisch gleich. Die Standardabweichungen bei gebundener Hochrechnung sind etwas kleiner als bei Designgewichtung. Jedoch sind die Unterschiede so minimal, dass auf eine weitere Ergebnisdiskussion verzichtet werden kann.

Beispiel 9: Durchschnittliche Arbeitsstunden (Designgewichtung)

```
*   sub: Bevölkerung am Hauptwohnsitz, Erwerbstätige
gen sub = ef505>=1 & ef505<=2 & ef504==1
*   ef141 Normalerw. geleist. Arbeitszeit (Std.) je Woche
recode ef141 (57=58) (60=62) (65=67) (70=72) (75=77) ///
             (80=82) (85=87) (90=93.5), gen(y) copyrest

svyset ef3 [pw=w] , strata(schicht) fpc(f) singleunit(certainty)
svy linearized, subpop(sub) : mean y,  over(ef32 westost)
  Survey: Mean estimation
  {...}
```

```
          Over: ef32 westost
      _subpop_1: [1] Männlich West
      _subpop_2: [1] Männlich Ost
      _subpop_3: [2] Weiblich West
      _subpop_4: [2] Weiblich Ost
------------------------------------------------------------
             |             Linearized
        Over |       Mean   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------
Y            |
   _subpop_1 |   40,74235   ,1629036      40,42303    41,06167
   _subpop_2 |   40,21037   ,2793079      39,66287    40,75786
   _subpop_3 |   30,86448   ,2160716      30,44094    31,28802
   _subpop_4 |   35,28901   ,3293919      34,64334    35,93468
------------------------------------------------------------
Note: 7 strata omitted because they contain no subpopulation members.

* Wald-Tests
* Arbeitszeit Männer West = Ost
test ([y]_subpop_1 = [y]_subpop_2)

         F(  1, 10585) =    2,71
              Prob > F =    0,1000

* Arbeitszeit Frauen West = Ost
test ([y]_subpop_3 = [y]_subpop_4)

         F(  1, 10585) =  126,16
              Prob > F =    0,0000
```

6.2 Poststratifikation

Analog zur Verhältnisschätzung ist der Populationsmittelwert bei gebundener Hochrechnung (Stata 2007b: 247; siehe dazu auch Abschnitt 4.4, S. 32)

$$\bar{y}^P = \hat{Y}^P / \hat{M}^P = \hat{Y}^P / \hat{M} \quad \text{mit}$$

$$\hat{Y}^P = \sum_{k=1}^{L_p} \frac{M_k}{\hat{M}_k} \sum_{j=1}^{m} I_{P_k}(j) w_j y_j \qquad \text{und}$$

$$\hat{M}^P = \sum_{k=1}^{L_P} M_k = M$$

Die Hilfsvariable für die Varianzschätzung ist

$$z_j\left(\bar{y}^P\right) = \frac{z_j\left(\hat{Y}^P\right)}{M} = \frac{1}{M} \sum_{k=1}^{L_P} I_{P_k}(j) \frac{M_k}{\hat{M}_k} \left(y_j - \frac{\hat{Y}_k}{\hat{M}_k} \right)$$

Beispiel 10: Durchschnittliche Arbeitsstunden mit gebundener Hochrechnung (Poststratifikation)

```
* Fortsetzung von Beispiel 9
* Anpassungsklassen (siehe Beispiel 5)
{...}
gen anpschicht=sub*(ef1*10+anp) // nur für Subpop
*    Populationsdaten pro Anpassungsschicht
egen M_k = total(ef750g*100*(ef505<=2)), by(anpschicht)
* Mittelwert Arbeitszeit mit gebundener Hochrechnung

svyset ef3 [pw = w], strata(schicht) fpc(f) ///
   single(certainty)  poststrata(anpschicht) postweight(M_k)
svy linearized, subpop(sub) : mean y,  over(ef32 westost)

  Survey: Mean estimation
  {...}
           Over: ef32 westost
       _subpop_1: [1] Männlich West
       _subpop_2: [1] Männlich Ost
       _subpop_3: [2] Weiblich West
       _subpop_4: [2] Weiblich Ost

  --------------------------------------------------------------
               |             Linearized
          Over |       Mean   Std. Err.     [95% Conf. Interval]
  -------------+------------------------------------------------
  y            |
     _subpop_1 |   40,68114   ,1609989      40,36555    40,99673
     _subpop_2 |   40,17563   ,2767858      39,63308    40,71818
     _subpop_3 |   30,84385   ,2169984       30,4185    31,26921
     _subpop_4 |   35,27467   ,3278229      34,63208    35,91726
  --------------------------------------------------------------
  Note: 7 strata omitted because they contain no subpopulation members.
```

7 Statistische Modelle

Wie die obigen Beispielanalysen zeigen, muss bei Populationsschätzungen für eine angemessene Schätzung der Varianz das Stichprobendesign berücksichtigt werden, da andernfalls bei der Annahme einer einfachen Zufallsauswahl die Stichprobenfehler des Mikrozensus gravierend unterschätzt werden. Das Survey-Kommando kann nicht nur für Punkt- bzw. Populationsschätzungen von Gesamt-, Verhältnis- und Mittelwerten, sondern auch für statistische Modelle eingesetzt werden.

Mit Punktschätzungen sollen Parameter der Gesamtheit unter Berücksichtigung des Stichprobendesigns geschätzt werden. Statistische Modelle dienen i. d. R. der hypothesengeleiteten Daten- und Informationsreduktion bzw. der Schätzung von Werten für abhängige Variablen unter der Bedingung beobachteter oder angenommener Werte unabhängiger Variablen. Vor dem Hintergrund dieser Zielsetzungen statistischer Modelle einerseits und der Modellannahmen oder –voraussetzungen andererseits (vgl. Deaton 1997: 63ff.; Rohwer und Pötter 2001: 111ff.) stellt sich die grundsätzliche Frage, ob die designbasierte Schätzung auch für Regressionskoeffizienten und die Standardfehler dieser Koeffizienten notwendig ist. Von entscheidender Bedeutung ist dabei die Feststellung, dass die Standardfehler der Modellparameter im Wesentlichen nicht vom Stichprobendesign, sondern vom Modell abhängen. Im Modell fehlende erklärende Variablen und die Modellierung nicht zutreffender funktionaler Beziehungen zwischen den Variablen (z. B. linearer statt kurvilinearer Zusammenhang) haben unabhängig davon, wie die Daten zustande gekommen sind, also unabhängig vom Stichprobendesign, einen erheblichen Einfluss auf die Standardfehler der Koeffizienten (Rohwer und Pötter 2001: 165). Aus statistischer und modelltheoretischer Sicht können in Bezug auf die Elemente des Stichprobendesigns die einzelnen Punkte zusammengefasst werden.

Schichtung. Sofern in den Schichten gleiche Auswahlwahrscheinlichkeiten vorliegen, ist die Schichtung aus modelltheoretischer Sicht nicht interessant. Falls es Annahmen gibt, dass die Schichtungsvariablen einen statistischen Einfluss auf eine interessierende Variable haben, sollten diese Schichtungsinformationen als erklärende Variablen in ein Modell aufgenommen werden.

Designgewichtung. Allerdings müssen in statistischen Modellen unterschiedliche Auswahlsätze z. B. in Form gewichteter Regressionen (WLS) berücksichtigt werden, da diese mit den erklärenden Variablen korreliert sein können und einen Einfluss auf den Standardfehler der Modellkoeffizienten haben (vgl. zusammenfassend Gabler 2006). Bis auf die mit einem

durchschnittlichen Auswahlsatz von rund 0,45 Prozent durchgeführten Substichproben zu den Ergänzungs- und Zusatzprogrammen, für deren Merkmale aufgrund unbekannter regional variabler Auswahlwahrscheinlichkeiten (siehe S. 13) keine angemessene Varianzschätzung möglich ist, liegt im Mikrozensus der einheitliche Auswahlsatz von einem Prozent zugrunde. In den anonymisierten Daten gilt der entsprechende Substichprobenauswahlsatz von 70 Prozent (SUF) bzw. 3,5 Prozent (CF). In einem statistischen Modell muss deshalb nicht mit dem für alle Beobachtungen konstanten Designgewicht gewichtet werden.

Poststratifikation. Die zur Anpassung der Mikrozensus-Ergebnisse an Randverteilungen der Population bereitstehenden Gewichtungsvariablen können für bestimmte deskriptive Analyseziele hilfreich sein (Deaton 1997: 71): „A weighted regression provides a consistent estimate of the population regression function (...). Of course, if we are trying to estimate behavioral models, and if those models are different in different parts of the population, the classic econometric argument is correct, and weighting is at best useless.“[22]

Die Populationsdaten aus der laufenden Bevölkerungsfortschreibung sind allerdings nicht fehlerfrei. Bei Verwendung der Hochrechnungsfaktoren werden die nach regionalen und groben demografischen Merkmalen ermittelten Untererfassungen des Mikrozensus im Vergleich zur Bevölkerungsfortschreibung auf alle anderen Merkmale übertragen. Ob die damit verbundene Annahme homogener, gruppenspezifischer Antwortwahrscheinlichkeiten zutrifft, ist jedoch fraglich. Für die meisten statistischen Modelle ist deshalb die gebundene Hochrechnung nicht zu empfehlen.

Klumpung. Aufgrund der Erfassung aller Einheiten eines ausgewählten Klumpens sind sich Einheiten eines Auswahlbezirks i. d. R. in ihren Merkmalen ähnlicher als Einheiten bei einfacher Zufallsauswahl. Die Klumpung kann deshalb zur Verletzung der den meisten statistischen Modellen zugrunde liegenden Annahme unabhängig identisch verteilter Residuen führen: „In practice, the design feature that usually has the largest effect on standard error is clustering, and the most serious problem with the conventional formulas is that they overstate precision by ignoring the dependence of observations within the same PSU“ (Deaton 1997: 71). Um die Verletzung der i. i. d.-Annahme zu umgehen, könnte man ein statistisches Modell für die Einheiten der Auswahlbezirke schätzen. In der Regel ist man aber an Personen oder Haushalten als Analyseeinheiten interessiert.

[22] Deaton 1997: 70: „This is the classic econometric argument against the weighted estimator: when the sectors are homogeneous, OLS is more efficient, and when they are not, both estimators [WLS und OLS; B.S.-N.] are inconsistent.”

Neben der Anwendung der Survey-Prozedur von Stata (2007a: 156-157) besteht eine weitere Lösung des Problems darin, eine sogenannte robuste Varianzschätzung vorzunehmen (Baum 2006: 133f; Deaton 1997: 74-78; Rogers 1993; Wooldridge 2002: 152, 401f.). Hierbei wird das Residuum aufgespalten in einen Fehlerterm für die Klumpen G_m, der zwischen den Klumpen unkorreliert ist, und einen Fehlerterm für die Untersuchungseinheiten. Siehe hierzu unten die blockweise Diagonalmatrix, in der Σ_m für die Kovarianzmatrix innerhalb eines Klumpens steht (Baum 2006: 135).

$$\Sigma_u = \begin{pmatrix} \Sigma_1 & 0 & & 0 \\ 0 & \ddots & & \\ & & \Sigma_m & \\ & & & \ddots \; 0 \\ 0 & & & 0 \; \Sigma_M \end{pmatrix}$$

Stata (2007c: 97; 2007d) verwendet bei der linearen Regression für die Schätzung der robusten Varianz die Formel:

$$\hat{V} = q_c \hat{\mathbf{V}} \left(\sum_{k=1}^{M} \mathbf{u}_k^{(G)\prime} \mathbf{u}_k^{(G)} \right) \hat{\mathbf{V}}$$

mit

$$q_c = M/(M-1)$$

$$\hat{\mathbf{V}} = (\mathbf{X'X})^{-1}$$

$$\mathbf{u}_k = \sum_k \mathbf{e}_j \mathbf{x}_j$$

q_c ist ein konstanter FPC-Faktor, $\hat{\mathbf{V}}$ die Kovarianzmatrix und $\mathbf{u}_k$ die Summe der mit dem Kovariatenvektor multiplizierten Schätzfehler $\boldsymbol{e}_j$ der Beobachtungen. Bei der robusten Varianzschätzung ändert sich der Regressionskoeffizient gegenüber einer regulären Schätzung nicht, sondern nur der Standardfehler des Koeffizienten. Die „Likelihood“ solcher Modelle wird als Pseudo-Likelihood bezeichnet, da sie nicht die Stichprobenverteilung wiedergibt.

Dies wird am Beispiel der Schätzung des logarithmierten monatlichen Nettoeinkommens mittels linearer Regression mit wenigen Variablen illustriert. Dazu werden die im Mikrozensus gruppiert vorliegenden Einkommensangaben auf die Klassenmitte rekodiert.[23] In Anlehnung an eine humankapitaltheoretische Einkommensfunktion werden für den allgemeinen Bil-

[23] Stata bietet für solche Daten eine Intervallregression (`intreg`) an. Die entsprechenden Ergebnisse unterscheiden sich in diesem Beispiel jedoch nicht wesentlich von denen der linearen Regression.

dungsabschluss und den beruflichen Abschluss typische Ausbildungsdauern in Jahren vergeben.[24] Als weitere erklärende Variablen werden nur das Geschlecht und die Staatsangehörigkeit ohne Interaktionen mit den Bildungsvariablen herangezogen. Da im Mikrozensus das Nettoeinkommen nicht nur Erwerbseinkommen, sondern alle Einkommensquellen umfasst, wird die Analyse auf abhängig beschäftigte Erwerbstätige ohne Auszubildende (ef127) mit überwiegendem Lebensunterhalt aus Erwerbstätigkeit (ef338, ef504) beschränkt. Außerdem werden Personen am Hauptwohnsitz (ef505) mit gültigen Bildungs- und Einkommensangaben ausgewählt.

Die verschiedenen Ergebnisse des Beispiels 11 sind in Tabelle 6 zusammengefasst. Zunächst wird die Einkommensfunktion mittels linearer Regression (1) unter der Annahme einer einfachen Zufallsstichprobe geschätzt. Der mit diesem stark vereinfachten Modell geschätzte Einkommenszuwachs beträgt pro zusätzlichem Ausbildungsjahr rund acht bis neun Prozent und ist für die (typischen) beruflichen Ausbildungsdauern (8,7 % = exp(0,0838)-1) etwas höher als für die allgemeinbildenden Abschlüsse (8,1 % = exp(0,0777)-1). Jeweils unter der Annahme durchschnittlicher Bildungsjahre betragen die geschätzten Einkommen ausländischer abhängig beschäftigter Männer im Vergleich zu deutschen Männern rund 95 Prozent (exp(-0,0510)), dagegen erzielen deutsche Frauen nur etwa 60 Prozent (exp(-0,4954)) des Einkommens deutscher Männer. Diese Ergebnisse sind allerdings aufgrund fehlender erklärender Variablen, wie z. B. die Arbeitsstunden, nur eingeschränkt interpretationsfähig.

Die der linearen Regression (1) zugrunde liegende Annahme unabhängig identisch verteilter Residuen lässt sich mit dem White-Test oder dem Breusch-Pagan-Test überprüfen. Beide testen die Nullhypothese, dass die Residualvarianzen homogen sind. Wie unten zu sehen, trifft dies offensichtlich nicht zu.

Die Standardfehler der Koeffizienten der Regression mit robuster Varianzschätzung (2) sind höher; und zwar bei der Variablen „Beruflicher Abschluss“ um 20 Prozent. Dies zeigt der Fehlspezifikationskoeffizient *MEFT* mit dem Verhältnis der designbasierten (2) zur OLS-Standardabweichung (1) an. Auch wenn es in diesem Beispiel nicht eintrifft, deutet sich damit an, dass aus der Verletzung der Verteilungsannahme andere Befunde der statistischen Signifikanz resultieren können.

[24] Bei der Lehrausbildung werden abweichend von anderen Studien volle drei Jahre ohne Anrechnung einer Erwerbszeit angesetzt.

Tabelle 6: Ergebnisse der Regressionen des logarithmierten monatlichen Nettoeinkommens auf Bildung, Geschlecht und Staatsangehörigkeit unter verschiedenen Modellannahmen (Beispiel 11)

Modell Koeffizient	Variable [Referenzkategorie]	1 Lineare Regression (OLS)	2 Robuste Varianzschätzung	3 SVY: Klumpung, Designgew.	4 SVY: Klumpung, Poststratifik.
b (Std.abw.)	Allgemeine Bildung in Jahren	0,0777 (0,0042)	0,0777 (0,0048)	0,0777 (0,0047)	0,0759 (0,0047)
MEFT			1,14	1,14	1,14
b (Std.abw.)	Beruflicher Abschluss in Jahren	0,0838 (0,0051)	0,0838 (0,0062)	0,0838 (0,0062)	0,0812 (0,0061)
MEFT			1,20	1,20	1,19
b (Std.abw.)	Geschlecht = weiblich	-0,4954 (0,0118)	-0,4954 (0,0124)	-0,4954 (0,0122)	-0,4941 (0,0120)
MEFT	[Ref. = männlich]		1,04	1,03	1,02
b (Std.abw.)	Staatsangehörigkeit = Ausländer	-0,0510 (0,0256)	-0,0510 (0,0257)	-0,0510 (0,0256)	-0,0556 (0,0243)
MEFT	[Ref. = Deutsche]		1,00	1,00	0,95
b (Std.abw.)	Konstante	6,4065 (0,0404)	6,4065 (0,0429)	6,4065 (0,0427)	6,4308 (0,0431)
MEFT			1,06	1,06	1,07
R^2		0,2750	0,2750	0,2750	0,2722

Schätzt man das Modell mit Statas Survey-Prozedur (3) unter Berücksichtigung der Schichtung, Klumpung und des (eigentlich nicht notwendigen; s. o.) Designgewichtes, ergeben sich gegenüber der robusten Varianzschätzung im Grunde keine nennenswerten Unterschiede. D. h., die im Modell 3 enthaltene Schichtung sowie die Annahme des Ziehens ohne Zurücklegen (FPC) sind praktisch unerheblich.[25]

Jedoch verändern sich im vierten Modell durch die Gewichtung mit dem Faktor der gebundenen Hochrechnung nicht nur die Koeffizienten zu Geschlecht und Staatsangehörigkeit, sondern auch die Regressionskoeffizienten der Bildungsvariablen. Im Vergleich zu den anderen drei Modellen fällt damit die „Bildungsrendite“ etwas geringer aus. Dies hängt überwiegend mit dem höheren Gewichtungsfaktor von Ausländern (siehe Tab. 3) zusammen, die zugleich ein unterdurchschnittliches Einkommen besitzen. Da unklar ist, ob die Anpassung an demografische Populationsdaten mit dem hier interessierenden Nettoeinkommen in einer (linearen) Beziehung steht, sollten diese Gewichte in der Regression nicht verwendet werden.

[25] Während mit der robusten Varianzschätzung die Klumpung nur für eine Stufe (hier z. B. Auswahlbezirk) berücksichtigt werden kann, können mit der Survey-Prozedur Schätzungen auch für mehrstufige Stichproben durchgeführt werden.

Beispiel 11: Regression des Monatsnettoeinkommens auf Bildung, Geschlecht und Staatsangehörigkeit

```
* Siehe Beispiel 10 zur Abgrenzung der Variablen schicht, f, w,
*  anpschicht und M_k
{...}
* (1) OLS (SRSWR)
xi: regress logv372m x287 x289 i.ef32 i.v52 if sub
{...} (siehe die Ergebnisse in Tabelle 6)

estat imtest, white

  White's test for Ho: homoskedasticity
           against Ha: unrestricted heteroskedasticity
           chi2(12)     =    171,20
           Prob > chi2  =    0,0000

estat hettest, iid

  Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
           Ho: Constant variance
           Variables: fitted values of logv372m
           chi2(1)      =    22,99
           Prob > chi2  =   0,0000

* (2) Robuste Varianzschätzung, Klumpung: Auswahlbezirk (SRSWR)
xi: regress logv372m x287 x289 i.ef32 i.v52 if sub, ///
              vce(cluster ef3)

* (3) SVY: Schichtung, Klumpung, Designgewichtung (SRSWOR)
svyset ef3 [pw=w] , strata(schicht) fpc(f) singleunit(certainty)
xi: svy: regress logv372m x287 x289 i.ef32 i.v52 if sub

* (4) SVY: Schichtung, Klumpung, gebundene Hochrechnung (SRSWOR)
svyset ef3 [pw=w] , strata(schicht) fpc(f) ///
       singleunit(certainty) poststrata(anpschicht) postweight(M_k)
xi: svy: regress logv372m x287 x289 i.ef32 i.v52 if sub
```

8 Zusammenfassung

Vernachlässigt man bei Analysen des Mikrozensus das Stichprobendesign, kann dies insbesondere aufgrund der Klumpenauswahl zu falsch berechneten Varianzen, Konfidenzintervallen und statistischen Tests führen. Da nicht nur Stata, sondern die meisten Statistikprogramme inzwischen Prozeduren zur Berücksichtigung komplexer Stichprobendesigns anbieten, ist es bei Punkt- bzw. Populationsschätzungen nicht mehr nötig, den Stichprobenfehler unter der Annahme einer einfachen Zufallsauswahl zu schätzen und für Korrekturen die von den statistischen Ämtern veröffentlichten Zuschlagsfaktoren zu verwenden. In statistischen Modellen kann mit der robusten Varianzschätzung die Klumpung im Mikrozensus sachgerecht berücksichtigt werden.

Veröffentlichte Ergebnisse der statistischen Ämter basieren auf einer gebundenen Hochrechnung, mit der die Fallzahlen des Mikrozensus an bekannte Populationsverteilungen angepasst werden. Bei Gewichtung der anonymisierten Daten mit den Hochrechnungsfaktoren werden vergleichbare Ergebnisse erzielt. Für die Varianzschätzung bei gebundener Hochrechnung kann im Survey-Kommando die Option „poststratification“ verwendet werden. In Bezug auf die in den anonymisierten Daten nur teilweise identifizierbaren Anpassungsschichten und weil sich die Gewichtungsfaktoren zwischen Gesamtpopulation und interessierendem Merkmal bzw. Bevölkerungsgruppe ggf. unterscheiden, ist die Prozedur nicht optimal, doch zeigen die Beispiele, wie Stata in der praktischen Analyse eingesetzt werden kann.

Das Campus File entstammt einer systematischen Zufallsauswahl. Um exemplarische Varianzschätzungen durchführen zu können, müssen stark vereinfachende Annahmen gemacht werden. Die Stata-Beispiele zum Campus File lassen sich ohne großen Aufwand auch auf das Scientific Use File des Mikrozensus bis 2004 übertragen, wobei die dafür getroffenen Annahmen zum Ziehungsverfahren wesentlich plausibler erscheinen. Für die Statistikprogramme R, SAS, und SPSS werden zu den Stata-Beispielen vergleichbare Programme bereit gestellt (siehe unter www.gesis.org/Dauerbeobachtung/ GML/Service/Mikrodaten-Tools/Varianz/varianz_ tools.htm). Welche Anpassungen bei der Analyse des Mikrozensus ab 2005 vorzunehmen sind, bleibt späteren Arbeiten vorbehalten.

Literatur

Afentakis, Anja, und Wolf Bihler, 2005: Das Hochrechnungsverfahren beim unterjährigen Mikrozensus ab 2005. Wirtschaft und Statistik (10): 1039-1048. URL: www.destatis.de/jetspeed/portal/cms/Sites/destatis/Internet/DE/Content/Publikationen/Querschnittsveroeffentlichungen/WirtschaftStatistik/Mikrozensus/Hochrechnungunterjaehrig,property=file.pdf; 08. 05. 2008.

Baum, Christopher F., 2006: An Introduction to Modern Econometrics Using Stata. College Station, TX: Stata Press.

Bihler, Wolf, 2008a: Das Hochrechnungsverfahren im Mikrozensus bis 2004. Vortrag zum Workshop "Stichprobendesign und Hochrechnungsverfahren im Mikrozensus“, 12.-13. Juni 2008, Mannheim. Wiesbaden: Statistisches Bundesamt.

Bihler, Wolf, 2008b: Parameterschätzungen aus den anonymisierten Mikrozensus-Files (II): gebundene Hochrechnung. Vortrag zum Workshop "Stichprobendesign und Hochrechnungsverfahren im Mikrozensus“, 12.-13. Juni 2008, Mannheim. Wiesbaden: Statistisches Bundesamt.

Breiholz, Holger, 2003: Ergebnisse des Mikrozensus 2002. Wirtschaft und Statistik (7): 601-610. URL: http://www.destatis.de/ jetspeed/ portal/ cms/ Sites/ destatis/ Internet/ DE/ Content/ Publikationen/ Querschnittsveroeffentlichungen/ WirtschaftStatistik/ Mikrozensus/ Mikrozensus2002, property = file.pdf.

Carlson, Barbara L., 1998: Software for Sample Survey Data. S. 4160-4167 in: Peter Armitage und Theodore Colton (Hg.): Encyclopedia of Biostatistics, Vol. 5. New York: Wiley.

Cochran, William G., 1972: Stichprobenverfahren. Berlin: de Gruyter.

Deaton, Angus, 1997: The Analysis of Household Surveys. Baltimore: The Johns Hopkins University Press.

Forschungsdatenzentrum der Statistischen Ämter des Bundes und der Länder (Website): Campus File Mikrozensus 2002 (Daten, Methodenbeschreibung, Schlüsselverzeichnis). URL: www.forschungsdatennetzwerk.de/bestand/mikrozensus/cf/2002/index.asp; 08. 05. 2008.

Gabler, Siegfried, und Horst Stenger, 2006: Systematische Primärauswahl mit Anwendungen in Wirtschafts- und Sozialstatistik. S. 37-50 in: Hans Wolfgang Brachinger, Alfred Hamerle, Ralf Münnich und Walter Schweitzer (Hg.): Wirtschaftsstatistik. Festschrift zum 65. Geburtstag von Professor Dr. Dr. h.c. mult. Eberhard Schaich. München: Vahlen.

Gabler, Siegfried, 2006: Gewichtungsprobleme in der Datenanalyse. S. 128-147 in: Andreas Diekmann (Hg.): Aktuelle Probleme der empirischen Sozialforschung. Kölner Zeitschrift für Soziologie und Sozialpsychologie, Sonderheft 44, 2004. Wiesbaden: VS Verlag für Sozialwissenschaften.

GESIS, German Microdata Lab (Website): Mikrodaten-Tools. Zur Berechnung des Stichprobenfehlers im Mikrozensus. URL: www.gesis.org/Dauerbeobachtung/GML/Service/Mikrodaten-Tools/Varianz/varianz_tools.htm.

Graubard, Barry I., und Edward L. Korn, 1996: Survey Inference for Subpopulations. American Journal of Epidemiology 44(1): 102-106.

Heidenreich, Hans-Joachim, 1994: Hochrechnung des Mikrozensus ab 1990. S. 112-123 in: Siegfried Gabler, Jürgen H.P. Hoffmeyer-Zlotnik und Dagmar Krebs (Hg.): Gewichtung in der Umfragepraxis. Opladen: Westdeutscher Verlag.

Heidenreich, Hans-Joachim, 1989: Erwerbstätigkeit im April 1988. Wirtschaft und Statistik (6): 327-339.

Herberger, Lothar, 1985: Aktualität und Genauigkeit der repräsentativen Statistik der Bevölkerung und des Erwerbslebens. Allgemeines Statistisches Archiv 69: 16-55.

Kohler, Ulrich, und Frauke Kreuter, 2008: Data Analysis Using Stata. College Station, TX: Stata Press (2. Auflage).

Kreuter, Frauke, und Richard Valliant, 2007: A survey on survey characteristics: What is done and can be done in Stata. The Stata Journal 7(1): 1-21.

Krug, Walter, Martin Nourney und Jürgen Schmidt, 2001: Wirtschafts und Sozialstatistik. Gewinnung von Daten. München: Oldenbourg (6. völlig neubearbearbeitete und erweiterte Auflage).

Meyer, Kurt, 1994: Zum Auswahlplan des Mikrozensus ab 1990. S. 106-111 in: Siegfried Gabler, Jürgen H.P. Hoffmeyer-Zlotnik und Dagmar Krebs (Hg.): Gewichtung in der Umfragepraxis. Opladen: Westdeutscher Verlag.

Opfermann, Heike, Claire Grobecker und Elle Krack-Roberg, 2006: Auswirkung der Bereinigung des Ausländerzentralregisters auf die amtliche Ausländerstatistik. Wirtschaft und Statistik (5): 480-494.

Rendtel, Ulrich, und Bernhard Schimpl-Neimanns, 2001: Die Berechnung der Varianz von Populationsschätzern im Scientific Use File des Mikrozensus ab 1996. ZUMA-Nachrichten 48: 85-116. URL: www.gesis.org/Publikationen/Zeitschriften/ZUMA_Nachrichten/documents/pdfs/48/zn48_10-bernhard.pdf.

Rogers, William, 1993: sg17: Regression standard errors in clustered samples. Stata Technical Bulletin 13:19-23.

Rohwer, Götz, und Ulrich Pötter, 2001: Grundzüge der sozialwissenschaftlichen Statistik. Weinheim: Juventa.

Särndal, Carl-Erik, Bengt Swensson und Jan Wretman, 1997: Model Assisted Survey Sampling. New York: Springer (korrigierte vierte Auflage).

Schimpl-Neimanns, Bernhard, und Ulrich Rendtel, 2001: SAS-, SPSS- und STATA-Programme zur Berechnung der Varianz von Populationsschätzern im Mikrozensus ab 1996. ZUMA-Methodenbericht 2001/04. URL: www.gesis.org/Publikationen/Berichte/ZUMA_Methodenberichte/documents/pdfs/tb01_04.pdf.

Stata, 2007a: Stata Survey Data Reference Manual. Release 10. College Station, TX: Stata Press.

Stata, 2007b: Stata Base Reference Manual. Volume 2, I-P. Release 10. College Station, TX: Stata Press.

Stata, 2007c: Stata Base Reference Manual. Volume 3, Q-Z. Release 10. College Station, TX: Stata Press.

Stata, 2007d: Comparison of standard errors for robust, cluster, and standard estimators. URL: www.stata.com/support/faqs/stat/cluster.html; 07.05.2008.

Statistisches Bundesamt, 2008: Qualitätsbericht Mikrozensus 2006. Wiesbaden. URL: http://www.destatis.de/jetspeed/portal/cms/Sites/destatis/Internet/DE/Content/Publikationen/Qualitaetsberichte/Mikrozensus/Mikrozensus2006,property=file.pdf.

Statistisches Bundesamt, o. J. [2008]: Konzept zur Anonymisierung des Mikrozensus 2002 zur Verwendung als CAMPUS File (CF). Wiesbaden. URL: www.forschungsdatennetzwerk.de/bestand/mikrozensus/cf/2002/fdz_mikrozensus_cf_2002_methodenbeschreibung.pdf.

Statistisches Bundesamt, 2006: Methodenpapier "Mikrozensus und Arbeitskräfte-erhebungen. Zur Problematik nicht-stichprobenbedingter Fehler". Wiesbaden. URL: www.destatis.de/jetspeed/portal/cms/Sites/destatis/Internet/DE/Content/Wissenschaftsforum/MethodenVerfahren/Mikrozensus/Veroeffentlichungen/PapierMikrozensusArbeitskraefteerhebung,property=file.pdf.

Statistisches Bundesamt, 2003a: Fachserie 1 / Reihe 4.1.1. Bevölkerung und Erwerbstätigkeit. Stand und Entwicklung der Erwerbstätigkeit. Ergebnisse des Mikrozensus 2002. Band 1: Allgemeine und methodische Erläuterungen. Wiesbaden. URL: https://www-ec.destatis.de, ..., 2010411027004.pdf.

Statistisches Bundesamt, 2003b: Fachserie 1 / Reihe 4.1.1. Bevölkerung und Erwerbstätigkeit. Stand und Entwicklung der Erwerbstätigkeit. Ergebnisse des Mikrozensus 2002. Bd. 2: Deutschland. Wiesbaden. URL: https://www-ec.destatis.de, …, 2010411027424.pdf.

UCLA Academic Technology Services, Statistical Consulting Group (Websites): “Stata Topics. Survey Data Analysis” (URL: www.ats.ucla.edu/stat/stata/topics/Survey.htm) und “Textbook Examples” (URL: www.ats.ucla.edu/stat/examples/default.htm; 08.05.2008).

Wells, Christine, 2007: Survey Data Analysis in Stata 10: Accessible and Comprehensive. Folien zum Vortrag beim 2007 West Coast Stata Users Group meeting. URL: www.stata.com/meeting/wcsug07/Wells_Stata10talk.pdf.

Wooldridge, Jeffrey M., 2002: Econometric Analysis of Cross Section and Panel Data. Cambridge, MA: MIT Press.

Anhang: Stata-Programme

Der Anhang enthält die Beispielprogramme zu diesem Methodenbericht. Sie zeigen, wie mit Stata (Version 10) Gesamt-, Verhältnis- und Mittelwerte bei freier Hochrechnung (Designgewichtung) sowie gebundener Hochrechnung (Poststratifikation, Redressment) mit dem Campus File Mikrozensus 2002 geschätzt werden können. In den hier verwendeten Daten sind keine fehlenden Werte mehr enthalten und weitere Modifikationen in enger Anlehnung an das Setup für das Scientific Use File Mikrozensus 2002 vorgenommen worden (siehe www.gesis.org/Dauerbeobachtung/GML/Daten/MZ/mz_2002/stata_setup02.htm). Insofern unterscheiden sie sich vom Campus File, wie es von den Forschungsdatenzentren bereitgestellt wird.

Für typische Schätzungen finden sich jeweils am Ende der Do-Files (nach „exit“) Hinweise auf Umsetzungen mit dem Scientific Use File Mikrozensus 2002 (siehe Beispiele 1, 3, 6 und 8). Vor Verwendung der Programme müssen die Verzeichnisse und ggf. die Dateinamen geändert werden.

Programmstatus: April 2009.

```
version 10
clear
capture log close
set more off
set memory 600m
set dp comma

* cd <Arbeitsverzeichnis>

log using Beisp_01.log, text replace

* Diese Datei: Beisp_01.do (20.06.2008)

* Beispiel 1: Designbasierte Schätzung des Gesamtwertes Zahl der
  Erwerbslosen

use ef1 ef3 ef4 ef7 ef30 ef504 ef505 ef506 ef507 ef708 ef712 ///
    using mz02cf_Beisp.dta, clear

* Designdefinition (A)
gen schicht = ef1*10 + ef712 // Bundesland, Gebäudegrößenklasse
gen f = 0.01 * 0.035  // Auswahlsatz MZ 1%, CF 3,5%
gen w = 1/f           // Designgewicht

* Interessierende Variable: ILO-Erwerbslos (ef504)
gen y = ef504==2

* Subpopulation: Bevölkerung am Hauptwohnsitz (ef505), 15+ Jahre (ef30)
gen sub = ef505>=1 & ef505<=2 & ef30>=15

* 1a) Ohne Ausschluss von PSUs mit nur einer Sekundäreinheit
*     Voreinstellung Option:single(missing)
* Definition Stichprobendesign (A)
* - Einstufige Klumpenauswahl: PSU = Auswahlbezirk (ef3)
* - Schichtung: schicht = Bundesland (ef1), Gebäudeschicht (ef712)
* - Endlichkeitskorrektur mit Auswahlsatz f = 0.01 * 0.035
* - Designgewicht: w (Inklusionswahrscheinlichkeiten MZ: 1% CF: 3,5%)
svyset ef3 [pw = w], strata(schicht) fpc(f)
* Schätzung des Gesamtwertes
svy linearized, subpop(sub) : total y

* 1b) Ausschluss von PSUs mit nur einer Sekundäreinheit
*     Option:singleunit(certainty)
svyset ef3 [pw = w], strata(schicht) fpc(f) single(certainty)
svy linearized, subpop(sub) : total y

* Problemfälle: PSUs mit nur einer Sekundäreinheit
svydescribe, single gen(single)
list ef1 ef712 ef3 ef505 ef506 y sub if single, nolab noobs

* 1c) Rekodierung von PSUs mit nur einer Sekundäreinheit in Pseudoschicht
gen pschicht = schicht
replace pschicht = 34 if single // Bremen & GU => Niedersachsen & GU
svyset ef3 [pw = w], strata(pschicht) fpc(f) single(missing)
svy linearized, subpop(sub) : total y

* 1d) Test alternativer Schätzungen

* (B) Einstufige Schätzung, Primäreinheit=Wohnung, Auswahlsatz f = 0,00035
gen whg = ef3*10+ef7
svyset whg [pw = w], strata(schicht) fpc(f) single(certainty)
```

```
svy linearized, subpop(sub) : total y

* (C) Zweistufige Schätzung,
*     1. Auswahlbezirk, Auswahlsatz f1 = 0,002376
*     2. Haushalt, Auswahlsatz      f2 = 0,00035 / f1 = 0,147
* gen f1 = 10707/(45058/0.01)
gen f2 = f / f1
replace w = 1/(f1*f2)
svyset ef3 [pw=w], strata(schicht) fpc(f1) vce(linearized)
      singleunit(certainty) || ef4, fpc(f2)
svy linearized, subpop(sub) : total y

* (D) Zweistufige Schätzung,
*     1. Auswahlbezirk, Auswahlsatz fMZ = 0,01
*     2. Haushalt, Auswahlsatz fCF = 0,035
gen f_MZ = 0.01
gen f_CF = 0.035
replace w = 1/(f_MZ * f_CF)
svyset ef3 [pw=w], strata(schicht) fpc(f_MZ) vce(linearized)
      singleunit(certainty) || ef4, fpc(f_CF)
svy linearized, subpop(sub) : total y


* === Erweiterte Schichtung ===
* + Gemeindegrößenklasse (ef708) als Proxy für Regionalschicht
gen schicht2 =  ef1*100 + ef708*10 + ef712

* (A') Einstufige Schätzung, PSU=Auswahlbezirk, f = 0.01 * 0.035
replace w = 1/f
svyset ef3 [pw = w], strata(schicht2) fpc(f) single(certainty)
quietly svydescribe, stage(1) single gen(s_A)
* tab s_A
svy linearized, subpop(sub) : total y

* (B') Einstufige Schätzung, PSU=Wohnung, Auswahlsatz f = 0,00035
svyset whg [pw = w], strata(schicht2) fpc(f) single(certainty)
quietly: svydescribe, stage(1) single gen(s_C)
svy linearized, subpop(sub) : total y

* (C') Zweistufige Schätzung, PSU=Auswahlbezirk, f1 = 0,002376
*                            SECU=Haushalt, f2 = 0,00035 / f1 = 0,147
replace w = 1/(f1*f2)
svyset ef3 [pw=w], strata(schicht2) fpc(f1) vce(linearized)
      singleunit(certainty) || ef4, fpc(f2)
quietly: svydescribe, stage(1) single gen(s_D1)
* tab s_D1
quietly: svydescribe, stage(2) single gen(s_D2)
* tab s_D2
svy linearized, subpop(sub) : total y

* (D') Zweistufige Schätzung, PSU=Auswahlbezirk, f_MZ= 0,01
*                            SECU=Haushalt, f_CF = 0,035
replace w = 1/(f_MZ * f_CF)
svyset ef3 [pw=w], strata(schicht2) fpc(f_MZ) vce(linearized)
      singleunit(certainty) || ef4, fpc(f_CF)
quietly: svydescribe, stage(1) single gen(s_E1)
* tab s_E1
quietly: svydescribe, stage(2) single gen(s_E2)
* tab s_E2
svy linearized, subpop(sub) : total y
```

```
exit


/* Beispiel 1 im Scientific Use File MZ 2002
* Einstufige Schätzung
* (...)
gen f1 = 0.01         // Auswahlsatz PSU:  1%
gen f2 = 0.70         // Auswahlsatz HH:  70%
gen w  = 1/(f1*f2)    // Designgewicht 1% * 70%
svyset ef3 [pw = w], strata(schicht) fpc(f1)
* svydescribe, stage(1) single gen(s1)
svy linearized, subpop(sub) : total y

* Zweistufige Schätzung
svyset ef3 [pw = w], strata(schicht) fpc(f1) || ef4, fpc(f2)
svydescribe, stage(2) single gen(s2)
svyset ef3 [pw = w], strata(schicht) fpc(f1) ///
    || ef4, fpc(f2) single(certainty)
svy linearized, subpop(sub) : total y
*/
```

```
version 10
clear
capture log close
set more off
set memory 200m
set dp comma

* cd <Arbeitsverzeichnis>

log using Beisp_02.log, text replace

* Diese Datei: Beisp_02.do (30.04.2009)

* Beispiel 2: Designbasierte Schätzung der Zahl der Erwerbstätigen (10
  PSUs)

use ef1 ef3 ef4 ef30 ef504 ef505 ef506 ef712  ///
    if ef3==0187 | ef3==2353 | ef3==3084 | ef3==4353 | ef3==6555 | ///
       ef3==6579 | ef3==7825 | ef3==8517 | ef3==9330 | ef3==9909   ///
    using mz02cf_Beisp.dta, replace

gen f = 0.01 * 0.035  // Auswahlsatz MZ 1%, CF 3,5%
gen w = 1/f           // Designgewicht

* Subpopulation Bev. in Privathaushalten am Hauptwohnsitz
*               im Alter von 15 Jahren und älter
gen sub = ef506==1 & ef505>=1 & ef505<=2 & ef30>=15

* Y-Variable: Erwerbstätige
gen y = ef504==1 /* Y-Variable ggf. mit sub multipliz. */

* Individualdaten
list ef3 y sub w, nolab sepby(ef3)

* Einstufige Klumpenausw., ohne Schichtung, Designgewichtung
svyset ef3 [pw = w], single(missing) fpc(f)
svy linearized, subpop(sub) : total y

/*  Berechnen der gewichteten PSU Totals und Anwendung der Stata-Formel im
    SVY-Manual, S. 151-152 sowie der ungewichteten PSU Totals für Anwendung
    der StBA-Formeln in Fachserie 1 Reihe 4.1.1, Mikrozensus 2002, S. 19 */
collapse (sum)  n_ghi=y , by(ef3)
gen y_hi = n_ghi * 1/(0.01 * 0.035)
* PSU- bzw. Haushaltsdaten
list , nolab sep(0)

* ==== Umsetzung StBA-Formeln ====
scalar f = 0.01 * 0.035  // Auswahlsatz
scalar m_h = _N
sum n_ghi, meanonly
scalar n_quer_gh = 1/_N * r(sum)
egen s2_gh = total((n_ghi - n_quer_gh)^2)
replace s2_gh = 1/(m_h-1) * s2_gh
scalar s2_gh = s2_gh[_N]

* (2) mit (n^2_g / f^2) multiplizieren
scalar var_y = (1-f) * m_h * s2_gh / (f*f) /* Varianz */
scalar s_y = var_y^.5                     /* Std.fehler */

* (1) mit Designgewichtung
gen total_y = sum(n_ghi*1/f)              /* Gesamtwert */
```

```
display as text "Total:  ", as res total_y[_N], ///
        _newline(1) as text "Std.Abw.:", as res s_y, ///
        _newline(1) as text "CV (%):    ", as res s_y*100/total_y[_N]

* ==== Umsetzung Stata-Formeln ====
scalar f_h = 0.01 * 0.035  // Auswahlsatz
scalar n_h = _N
sum y_hi, meanonly
scalar y_quer_h = 1/_N * r(sum)
egen v_y = total((y_hi - y_quer_h)^2)
replace v_y = (1-f_h) * n_h/(n_h-1) * v_y  /* Varianz */
scalar se_y = v_y[_N]^.5                   /* Std.fehler */
gen t_y = sum(y_hi)                        /* Gesamtwert */

display as text "Total:  ", as res t_y[_N], ///
        _newline(1) as text "Std.Abw.:", as res se_y, ///
        _newline(1) as text "CV (%):    ", as res se_y*100/t_y[_N]

exit
```

```
version 10
clear
capture log close
set more off
set memory 200m
set dp comma

* cd <Arbeitsverzeichnis>

log using Beisp_03.log, text replace

* Diese Datei: Beisp_03.do (22.05.2008)

* Beispiel 3a: Designeffekte im Campus File

use ef1 ef3 ef30 ef504 ef505 ef506 ef507 ef712 ///
    using mz02cf_Beisp.dta, clear

* Schichtung: Bundesland (ef1), Gebäudegrößenklasse (ef712)
gen schicht = ef1*10 + ef712

gen f = 0.01 * 0.035  // Auswahlsatz MZ 1%, CF 3,5%
gen w = 1/f           // Designgewicht

gen y = ef504==2 // Interessierende Variable: ILO-Erwerbslos (ef504)

* Subpopulation: Bevölkerung am Hauptwohnsitz (ef505), 15+ Jahre (ef30)
gen sub = ef505>=1 & ef505<=2 & ef30>=15

* Beispiel 3a: Designeffekte im Campus File
svyset ef3 [pw=w] , strata(schicht) fpc(f) singleunit(certainty)
quietly: svy linearized, subpop(sub) : total y
estat effects, deff deft meff meft
matrix v_d = e(V)   /* Designbasierte Varianz */
svmat  v_d

* SRSWOR: Einf. Zufallsstichprobe, Ziehen ohne Zurücklegen
svyset _n [pw=w], fpc(f)
svy linearized, subpop(sub) : total y
matrix v_srswor = e(V)
svmat  v_srswor

* SRSWR: Einfache Zufallsstichprobe, Ziehen mit Zurücklegen
svyset _n [pw=w]
svy linearized, subpop(sub) : total y
matrix v_srswr = e(V)
svmat  v_srswr

* SRSWR wie für Berechnung MEFF und MEFT
total y [pw=w] if sub
matrix v_msp = e(V)
svmat  v_msp

gen DEFF = v_d/v_srswor
gen DEFT = (v_d/v_srswr)^.5
gen MEFF = v_d/v_msp
gen MEFT = (v_d/v_msp)^.5
list DEFF DEFT MEFF MEFT in 1/1, noobs

exit
```

```
/*
* Beispiel 3b: Designeffekte im Scientific Use File MZ 2002
* Stata-Programm wie Beispiel 1 und Beispiel 3a, Y = Erwerbslos
* im SUF abweichende Stichprobendefinitionen
gen f = 0.01
gen w = 1/(0.01 * 0.70)  /* Designgewicht */
{...}
estat effects, deff deft meff meft

    |          Linearized
    |    Total  Std. Err.    DEFF     DEFT     MEFF     MEFT
----+------------------------------------------------------------
  y |  2930572  23156,14  1,34465  1,15552  1,34622  1,16027
*/
```

```
version 10
clear
capture log close
set more off
set memory 200m
set dp comma

* cd <Arbeitsverzeichnis>

log using Beisp_04.log, text replace

* Diese Datei: Beisp_04.do (20.06.2008)

* Beispiel 4: Gruppenvergleiche

use ef1 ef3 ef4 ef30 ef32 ef504 ef505 ef506 ef708 ef712 ///
    using mz02cf_Beisp.dta, clear

* Schichtung: Bundesland (ef1), Gebäudegrößenklasse (ef712)
gen schicht = ef1*10 + ef712
gen f = 0.01 * 0.035
gen w = 1/f
svyset ef3 [pw=w] , strata(schicht) fpc(f) singleunit(certainty)

* Interessierende Variable: ILO-Erwerbslos (ef504)
recode ef504 (2=1 "Erwerbslos") (*=0 "Sonst"), gen(y)
label var y "ILO-Erwerbsstatus" // nur bei Verwendung „sub" !

* Subpopulation: Bevölkerung am Hauptwohnsitz (ef505), 15-65 Jahre (ef30),
  Erwerbspers.
gen sub = ef505>=1 & ef505<=2 & ef30>=15 & ef30<=65 & ef504<=2

* West-/Ostdeutschland
recode ef1 (1/11=1 "West") (*=2 "Ost"), gen(westost)
replace westost = 2 if ef708==9  /* Ost-Berlin */
label var westost "West-/Ostdeutschland"

* Schätzung des Gesamtwertes Zahl der Erwerbslosen
svy linearized, subpop(sub) : total y

* ... in West-/Ostdeutschland
svy linearized, subpop(sub) : total y, over(westost)


* ... Fallzahltabelle
svy linearized, subpop(sub) : tabulate y westost, count ///
                format(%8.0f) cellwidth(10) stubwidth(20)

* ... mit Spalten-% und Unabhängigkeitstest
svy linearized, subpop(sub): tabulate y westost, col ///

    deft pearson  format(%5.3f) cellwidth(15) stubwidth(20)
exit
```

```
version 10
clear
capture log close
set more off
set memory 200m
set dp comma

cd F:\bsn\Workshops\WS2008\Beispiele

log using Beisp_05.log, text replace

* Diese Datei: Beisp_05.do (22.05.2008)

* Beispiel 5: Gesamtwerte mit gebundener Hochrechnung

use ef1 ef3 ef30 ef32 ef52 ef127 ef504 ef505 ef712 ef750g ///
    using mz02cf_Beisp.dta, replace

* Variablenabgenzungen analog zu Beispiel 1b und 3a
gen schicht = ef1*10 + ef712   // Bundesland + Gebäudegrößenklasse
gen f = 0.01 * 0.035           // Auswahlsatz MZ 1%, CF 3,5%
gen w = 1/f                    // Designgewicht
gen sub = ef505<=2 & ef30>=15  // Bev. am Hauptwohnsitz, 15+ Jahre
gen y = ef504==2 * sub         // ILO-Erwerbslos & Subpopulation

* Gesamtwert designgewichtet
svyset ef3 [pw = w], strata(schicht) fpc(f) ///
                     single(certainty)
svy linearized, subpop(sub) : total y , noheader
* Ergebnisse speichern
matrix t_d = e(b)   // Gesamtwert
svmat  t_d          //   Matrix -> Variable
matrix v_d = e(V)   // Varianz
svmat  v_d          //   Matrix -> Variable
disp "CV (%) = " (v_d^.5 * 100)/t_d   // Variationskoeffizient

* Vorbereitung Poststratifikation
* 1a) Anpassung: Geschl. * Staatsangeh., Sold., Wehrpfl.
gen anp=(1*(ef32==1 & ef52==1 & ef127~=9 & ef127~=10)) ///
      + (2*(ef32==2 & ef52==1)) ///
      + (3*(ef32==1 & ef52~=1)) ///
      + (4*(ef32==2 & ef52~=1)) ///
      + (5*(ef32==1 & ef52==1 & ef127==9)) ///
      + (6*(ef32==1 & ef52==1 & ef127==10))
lab var anp "Anpassungsklassen"
lab def anp 1 "Deutsche Maenner"  2 "Deutsche Frauen" ///
              3 "Ausl. Maenner"     4 "Ausl. Frauen" ///
              5 "Zeit-/Berufssold." 6 "Wehrpflichtige"
label val anp anp
* 1b) Proxy Anpassungsschicht
gen anpschicht=sub*(ef1*10+anp) /* nur für Subpop */
lab var anpschicht "Anpassungschicht - Proxy"

* 2)  Populationsdaten pro Anpassungsschicht
* egen  M_k = total(ef750g*100*(ef505<=2)), by(anpschicht)
* in zwei Schritten:
gen w_anp   = ef750g*100*(ef505<=2)   // Bevölkerung am Hauptwohnsitz
egen    M_k = total(w_anp), by(anpschicht)

* Total gebundene Hochrechnung
svyset ef3 [pw = w], strata(schicht) fpc(f) single(certainty) ///
```

```
                    poststrata(anpschicht) postweight(M_k)
svy linearized, subpop(sub) : total y
matrix t_p = e(b)
svmat  t_p
matrix v_p = e(V)
svmat  v_p
disp "CV (%) = " (v_p^.5 * 100)/t_p

* 3) Umsetzung der Stata-Formeln
egen M_Dach_k = total(w*sub), by(anpschicht)
gen w_post = M_k/M_Dach_k*w                 /* w*_j  */
replace w_post = 0 if sub==0
gen y_post = w_post*y
* Gesamtwert y_post: poststratifiziert
* <=> Abweichung z.B. zu: tab y if sub [iw=w_anp]
table y if y & sub, c(sum y_post sum w_anp)

* V(Y_post) für Hilfsmerkmal (score variable) z
egen Y_Dach_k = total(w*y*sub), by(anpschicht)
gen z = sub*(M_k/M_Dach_k) * (y - (Y_Dach_k/M_Dach_k))
replace z=0 if z==.
svyset ef3 [pw = w], strata(schicht) fpc(f) ///
                    single(certainty)
* "Total" irrelevant, nur S.E. interessiert
quietly: svy linearized, subpop(sub) : total z, noheader
matrix v_z = e(V)
svmat  v_z
disp "s.e. (z) = " v_z^.5

/* ------------------------------------------------------------------------
Problem:  Gesamtwert der gebundenen Hochrechnung stimmt nicht mit
          gewichteter Tabellierung überein (z.B. tab y if sub [iw=w_anp]
          bzw. (in 100): (...) [iw=ef750g]), da sich Hochrechnungsfaktoren
          insgesamt bzw. der Subpopulation von den Hochrechnungsfaktoren
          für y unterscheiden.
Lösung:   Direkte Verwendung des Korrekturfaktors ("Soll durch Ist") für y
          und Übertragung auf die Sub-/Population.
Ergebnis: Gesamtwert der SVY-Auswertung mit gebundener Hochrechnung
          entspricht einfacher gewichteter Auswertung.
          Standardfehler können sich je nach Gewichtung unterscheiden.
          Allerdings ändern sich Schätzungen der Population und
          Subpopulation ("Population size" u. "Subpop. size").
------------------------------------------------------------------------
*/
egen wy = total(ef750g*0.035*y), by(anpschicht)
egen ny = total(y), by(anpschicht)
gen  g_y = wy/ny // Korrekturfaktor der y-Werte
recode g_y (.=0)
gen w_y = g_y*w  // Endgewicht = Korrekturfaktor * Designgewicht
egen M_k2 = total(w_y), by(anpschicht)
svyset ef3 [pw=w], strata(schicht) fpc(f) single(certainty) ///
                   poststrata(anpschicht) postweight(M_k2)
svy linearized, subpop(sub) : total y
matrix t_k = e(b)
svmat  t_k
matrix v_k = e(V)
svmat  v_k
disp "CV (%) = " (v_k^.5 * 100)/t_k

exit
```

```
version 10
clear
capture log close
set more off
set memory 200m
set dp comma

* cd <Arbeitsverzeichnis>

log using Beisp_06.log, text replace

* Diese Datei: Beisp_06.do (22.05.2008)

* Beispiel 6: Gesamtwerte mit gebundener Hochrechnung
  (Regressionsschätzung)

use ef1 ef3 ef4 ef30 ef32 ef52 ef127 ef504 ef505 ef712 ef750g ///
    using mz02cf_Beisp.dta, replace

gen schicht = ef1*10 + ef712
gen f = 0.01 * 0.035  // Auswahlsatz MZ 1%, CF 3,5%
gen d = 1/f           // Designgewicht
gen sub = ef505<=2 & ef30>=15 /* Bev. am Hauptwohnsitz, 15+ Jahre */
gen y = ef504==2 * sub        /* ILO-Erwerbslos & Subpopulation */

* 1) Proxy Anpassungsschicht
gen anp=(1*(ef32==1 & ef52==1 & ef127~=9 & ef127~=10)) /// Deutsche Männer
      + (2*(ef32==2 & ef52==1)) /// Deutsche Frauen
      + (3*(ef32==1 & ef52~=1)) /// Ausländische Männer
      + (4*(ef32==2 & ef52~=1)) /// Ausländische Frauen
      + (5*(ef32==1 & ef52==1 & ef127==9)) /// Zeit-/Berufssoldaten
      + (6*(ef32==1 & ef52==1 & ef127==10)) // Wehrpflichtige
gen anpschicht=sub*(ef1*10+anp) /* nur für Subpop */
lab var anpschicht "Anpassungschicht - Proxy"

* 2) Berechnen des Regressionskoeffizienten mit Regression y  "Erwerbslos"
*    auf X "Hilfsmerkmale" bzw. "Anpassungsschicht"
*       xi: regress y i.anpschicht if sub,
*       predict B_Dach if e(sample), xb
*       replace B_Dach = 0 if B_Dach==.
* Da Anpassungsschichten disjunkt: B-Dach_g = y-quer_s_g = y_Dach.
* Einfacher:
egen B_Dach = mean(y), by(anpschicht)

* [A] Gewichtung analog Beispiel 5 (w_anp)
*  3) Berechnung des gewichteten Gesamtwertes über die Anpassungsschichten
*       ef750g rechnet auf 100 in Population hoch
egen    M_k   = total(ef750g*100*(ef505<=2)), by(anpschicht) //
  Populationswerte geb. Hochrechnung
egen M_Dach_k = total(d*(ef505<=2)), by(anpschicht) // Populationswerte
  designgewichtet
gen g1 = M_k/M_Dach_k     // Korrekturfaktor g_k
replace g1 = 0 if sub==0
gen  w1 = d*g1            // Endgewicht w_k
egen t_y1 = total(y*w1)
disp "Gewichteter Gesamtwert t_y1 = " t_y1

*  4) Varianzschätzung mit Hilfsvariable u
gen u1 = g1*(y-B_Dach)
svyset ef3 [pw = d], strata(schicht) fpc(f) single(certainty)
*  "total" irrelevant
```

```
svy linearized, subpop(sub) : total u1 , noheader

* [B] Gewichtung analog Beispiel 5 (w_y)
egen wy = total(ef750g*0.035*y), by(anpschicht)
egen ny = total(y), by(anpschicht)
gen  g2 = wy/ny // Korrekturfaktor der y-Werte
recode g2 (.=0)
gen w2 = g2*d  // Endgewicht = Korrekturfaktor * Designgewicht
egen t_y2 = total(y*w2)
disp "Gewichteter Gesamtwert t_y2 = " t_y2
gen u2    = g2*(y-B_Dach)
svyset ef3 [pw = d], strata(schicht) fpc(f) single(certainty)
svy linearized, subpop(sub) : total u2 , noheader

* [C] Gewichtung mit ef750 direkt (ohne Normierung auf 1%-MZ) analog zu
*     früheren Programmen (Rendtel/Schimpl-Neimanns 2001); z.B. varmz_a.do
gen  g3 = ef750g*0.035*sub
gen  w3 = d*g3
egen t_y3 = total(y*w3)
disp "Gewichteter Gesamtwert t_y3 = " t_y3
gen u3    = g3*(y-B_Dach)
svyset ef3 [pw = d], strata(schicht) fpc(f) single(certainty)
svy linearized, subpop(sub) : total u3 , noheader

exit

/* ----------------------------------------------------------------
* Beispiel 6 [C] mit Scientific Use File MZ 2002
* - im SUF abweichende Definitionen -
gen f = 0.01
gen d = 1/(0.01 * 0.70)   // Designgewicht
gen g = ef750*sub         // Korrekturfaktor
gen w = d*g               // Endgewicht
egen t_y = total(y*w)       // Gewichteter Gesamtwert
gen u    = g*(y-B_Dach)   // Hilfsvariable v. Varianzschätzung
svyset ef3 [pw = d], strata(schicht) fpc(f) single(certainty)
svy linearized, subpop(sub) : total u , noheader
* SUF-Ergebnisse:  Gesamtwert = 3.492.013    S.E. = 26.685
---------------------------------------------------------------- */
```

```
version 10
clear
capture log close
set more off
set memory 200m
set dp comma

* cd <Arbeitsverzeichnis>

log using Beisp_07.log, text replace

* Diese Datei: Beisp_07.do (22.05.2008)

* Beispiel 7: Erwerbslosenquote mit Designgewichtung

use ef1 ef3 ef4 ef30 ef32 ef52 ef127 ef504 ef505 ef708 ef712 ef750g ///
    using mz02cf_Beisp.dta, replace

* Variablenabgenzungen analog zu Beispiel 4
gen schicht = ef1*10 + ef712
gen f = 0.01 * 0.035  // Auswahlsatz MZ 1%, CF 3,5%
gen w = 1/f           // Designgewicht

recode ef1 (1/11=1 "West") (*=2 "Ost"), gen(westost)
replace westost = 2 if ef708==9
recode ef30 (0/14=0 "0-14")   (15/24=1 "15-24") ///
            (25/54=2 "25-54") (55/65=3 "55-65") ///
            (66/99=4 "66+"), gen(alter)

* sub: Bevölkerung am Hauptwohnsitz, 15-65 Jahre, Erwerbspersonen
gen sub = ef505>=1 & ef505<=2 & ef30>=15 & ef30<=65 & ef504>=1 & ef504<=2
gen y = (ef504==2) * sub            //Erwerbslose
gen x = (ef504>=1 & ef504<=2) * sub // Erwerbspersonen

* Erwerbslosenquote nach Altersgruppen
svyset ef3 [pw=w] , strata(schicht) fpc(f) singleunit(certainty)
quietly: svy linearized, subpop(sub) : ///
               ratio (Erwerbslosenquote: y/x), over(alter)
estat effects, deft
* Variationskoeffizienten (cv)
matrix ta = e(b)
matrix va = e(V)
matrix cva = sqrt(va[1,1])/ta[1,1] , sqrt(va[2,2])/ta[1,2], ///
             sqrt(va[3,3])/ta[1,3]
matrix list cva

* Erwerbslosenquote nach Region und Altersgruppen
quietly: svy linearized, subpop(sub) : ///
         ratio (Erwerbslosenquote: y/x), over(westost alter)
estat effects, deft

* Wald-Tests
* Erwerbslosenquote 15-24-Jähriger: West = Ost ?
test [Erwerbslosenquote]_subpop_1 = [Erwerbslosenquote]_subpop_4

* Erwerbslosenquote 55-65-Jähriger: West = Ost ?
test [Erwerbslosenquote]_subpop_3 = [Erwerbslosenquote]_subpop_6

exit
```

```
version 10
clear
capture log close
set more off
set memory 200m
set dp comma

* cd <Arbeitsverzeichnis>

log using Beisp_08.log, text replace

* Diese Datei: Beisp_08.do (22.05.2008)

* Beispiel 8: Erwerbslosenquote mit gebundener Hochrechnung
  (Poststratifikation)

use ef1 ef3 ef4 ef30 ef32 ef52 ef127 ef504 ef505 ef708 ef712 ef750g ///
    using mz02cf_Beisp.dta, replace

* Variablenabgenzungen analog zu Beispiel 4
gen schicht = ef1*10 + ef712
gen f = 0.01 * 0.035  // Auswahlsatz MZ 1%, CF 3,5%
gen w = 1/f           // Designgewicht

* sub: Bevölkerung am Hauptwohnsitz, 15-65 Jahre, Erwerbspersonen
gen sub = ef505>=1 & ef505<=2 & ef30>=15 & ef30<=65 & ef504>=1 & ef504<=2
gen y = (ef504==2) * sub            // Erwerbslose
gen x = (ef504>=1 & ef504<=2) * sub // Erwerbspersonen

recode ef1 (1/11=1 "West") (*=2 "Ost"), gen(westost)
replace westost = 2 if ef708==9
recode ef30 (0/14=0 "0-14")   (15/24=1 "15-24") ///
            (25/54=2 "25-54") (55/65=3 "55-65") ///
            (66/99=4 "66+"), gen(alter)

* Proxy Anpassungsschicht
gen anp=(1*(ef32==1 & ef52==1 & ef127~=9 & ef127~=10)) /// Deutsche Männer
      + (2*(ef32==2 & ef52==1)) /// Deutsche Frauen
      + (3*(ef32==1 & ef52~=1)) /// Ausländische Männer
      + (4*(ef32==2 & ef52~=1)) /// Ausländische Frauen
      + (5*(ef32==1 & ef52==1 & ef127==9)) /// Zeit-/Berufssoldaten
      + (6*(ef32==1 & ef52==1 & ef127==10)) // Wehrpflichtige
gen anpschicht=sub*(ef1*10+anp) /* nur für Subpop */

* Populationsdaten pro Anpassungsschicht (wie [A] in Beisp_06.do)
egen M_k = total(ef750g*100*(ef505<=2)), by(anpschicht)

* Erwerbslosenquote nach Altersgruppen
svyset ef3 [pw = w], strata(schicht) fpc(f) single(certainty) ///
                     poststrata(anpschicht) postweight(M_k)
svy linearized, subpop(sub) : ratio (Erwerbslosenquote: y/x), ///
                                    over(alter)

* Variationskoeffizienten (cv)
matrix tap = e(b)
matrix vap = e(V)
matrix cvap = sqrt(vap[1,1]) / tap[1,1] , ///
              sqrt(vap[2,2]) / tap[1,2], sqrt(vap[3,3]) / tap[1,3]
matrix list cvap

* Erwerbslosenquote nach Region und Altersgruppen
```

```
svy linearized, subpop(sub) : ratio (Erwerbslosenquote: y/x), ///
                                    over(westost alter)

* Wald-Tests
* Erwerbslosenquote 15-24-Jähriger: West = Ost ?
test [Erwerbslosenquote]_subpop_1 = [Erwerbslosenquote]_subpop_4

* Erwerbslosenquote 55-65-Jähriger: West = Ost ?
test [Erwerbslosenquote]_subpop_3 = [Erwerbslosenquote]_subpop_6

exit

/*
* Beispiel 8: Verhältniswerte bei gebundener Hochrechnung im SUF MZ 2002
* - abweichende Definitionen  -
gen f = 0.01
gen w = 1/(0.01 * 0.70)   // Designgewicht
egen M_k = total(ef750*100/0.7*(ef505<=2)), by(anpschicht)
(...)
svyset ef3 [pw = w], strata(schicht) fpc(f) single(certainty) ///
                     poststrata(anpschicht) postweight(M_k)
svy linearized, subpop(sub) : ratio (Erwerbslosenquote: y/x)

-------------------------------------------------------------
             |             Linearized
             |      Ratio   Std. Err.     [95% Conf. Interval]
-------------+-----------------------------------------------
Erwerbslos~e |   ,0879048   ,0006554      ,0866203    ,0891894
-------------------------------------------------------------
*/
```

```
version 10
clear
capture log close
set more off
set memory 200m
set dp comma

* cd <Arbeitsverzeichnis>

log using Beisp_09.log, text replace

* Diese Datei: Beisp_09.do (22.05.2008)

* Beispiel 9: Durchschnittliche Arbeitsstunden (Designgewichtung)

use ef1 ef3 ef32 ef141 ef504 ef505 ef708 ef712 ef750g ///
    using mz02cf_Beisp.dta, replace

* Variablenabgenzungen analog zu Beispiel 4
gen schicht = ef1*10 + ef712
gen f = 0.01 * 0.035  // Auswahlsatz MZ 1%, CF 3,5%
gen w = 1/f

*   westost: West-/Ostdeutschland
recode ef1 (1/11=1 "West") (*=2 "Ost"), gen(westost)
replace westost = 2 if ef708==9

* sub: Bevölkerung am Hauptwohnsitz, Erwerbstätige
gen sub = ef505<=2 & ef504==1

* ef141 Normalerw. geleist. Arbeitszeit (Std.) je Woche
recode ef141 (57=58) (60=62) (65=67) (70=72) (75=77) ///
             (80=82) (85=87) (90=93.5) (.=0), gen(y) copyrest

svyset ef3 [pw=w] , strata(schicht) fpc(f) singleunit(certainty)
svy linearized, subpop(sub) : mean y,  over(ef32 westost)

* Wald-Tests
* Arbeitszeit Männer West = Ost ?
test ([y]_subpop_1 = [y]_subpop_2)
* Arbeitszeit Frauen West = Ost ?
test ([y]_subpop_3 = [y]_subpop_4)

exit
```

```
version 10
clear
capture log close
set more off
set memory 200m
set dp comma

* cd <Arbeitsverzeichnis>

log using Beisp_10.log, text replace

* Diese Datei: Beisp_10.do (22.05.2008)

* Beispiel 10: Durchschnittliche Arbeitsstunden mit gebundener Hochrechnung
(Poststratifikation)


use ef1 ef3 ef32 ef52 ef127 ef141 ef504 ef505 ef708 ef712 ef750g ///
    using mz02cf_Beisp.dta, replace

* Variablenabgenzungen analog zu Beispiel 4
gen schicht = ef1*10 + ef712
gen f = 0.01 * 0.035  // Auswahlsatz MZ 1%, CF 3,5%
gen w = 1/f           // Designgewicht

* sub: Bevölkerung am Hauptwohnsitz, Erwerbstätige
gen sub = ef505<=2 & ef504==1

* ef141 Normalerw. geleist. Arbeitszeit (Std.) je Woche
recode ef141 (57=58) (60=62) (65=67) (70=72) (75=77) ///
             (80=82) (85=87) (90=93.5) (.=0), gen(y) copyrest

*   westost: West-/Ostdeutschland
recode ef1 (1/11=1 "West") (*=2 "Ost"), gen(westost)
replace westost = 2 if ef708==9

* Proxy Anpassungsschicht
gen anp=(1*(ef32==1 & ef52==1 & ef127~=9 & ef127~=10)) /// Deutsche Männer
      + (2*(ef32==2 & ef52==1)) /// Deutsche Frauen
      + (3*(ef32==1 & ef52~=1)) /// Ausländische Männer
      + (4*(ef32==2 & ef52~=1)) /// Ausländische Frauen
      + (5*(ef32==1 & ef52==1 & ef127==9)) /// Zeit-/Berufssoldaten
      + (6*(ef32==1 & ef52==1 & ef127==10)) // Wehrpflichtige
gen anpschicht=sub*(ef1*10+anp) /* nur für Subpop */

* Populationsdaten pro Anpassungsschicht (wie [A] in Beisp_06.do)
egen M_k = total(ef750g*100*(ef505<=2)), by(anpschicht)

* Mittelwert Arbeitszeit mit gebundener Hochrechnung
svyset ef3 [pw = w], strata(schicht) fpc(f) ///
   single(certainty)  poststrata(anpschicht) postweight(M_k)
svy linearized, subpop(sub) : mean y,  over(ef32 westost)

exit
```

```
version 10
clear
capture log close
set more off
set memory 200m
set dp comma

* cd <Arbeitsverzeichnis>

log using Beisp_11.log, text replace

* Diese Datei: Beisp_11.do (22.05.2008)

* Beispiel 11: Regression des Monatsnettoeinkommens auf Bildung, Geschlecht
  und Staatsangehörigkeit

use ef1 ef3 ef4 ef30 ef32 ef52 ef127 ef286 ef287 ef288 ef289 ///
    ef338 ef372 ef504 ef505 ef708 ef712 ef750g ///
   using mz02cf_Beisp.dta, replace

* Allgemeiner Bildungsabschluss - typische Ausbildungsdauer
recode ef287 (1=9 "[9] Hauptschule") ///
               (2 3=10 "[10] Realschule, POS") ///
               (4=12 "[12] FHR") ///
               (5=13 "[13] Abitur") ///
               (0=8 "[8] ohne Abschluss") ///
               (9=.a "k.A., Entfällt") ///
               (.a=.a "[.a] k.A.Entf"), gen(x287)
replace x287 = .a if (ef286==0 | ef286==9)

* Beruflicher Abschluss - typische Ausbildungsdauer
recode ef289 (0=0 "[0] ohne Abschluss") ///
             (1 2=1 "[1] Anlernausb., BVJ") ///
             (3 7 8=3 "[3] Lehre, FH, Verw.FH") ///
             (4 5 6=2 "[2] BFS, Meister, FS DDR") ///
             (9=5 "[5] Uni") ///
             (10=7 "[7] Prom.")  ///
             (0 99=.a "[.a] k.A., Entfällt"), gen(x289)
replace x289 = .a if ef288==0 | ef288==9

* Rekodierung Nettoeinkommen - Klassenmitte bzw.
*   obere Randklasse  = 150% * Untergrenze
recode ef372 (1=75) (2=225) (3=400) (4=600) (5=800) ///
  (6=1000) (7=1200) (8=1400) (9=1600) (10=1850) (11=2150) ///
  (12=2450) (13=2750) (14=3050) (15=3400) (16=3800) (17=4250) ///
  (18=4750) (19=5250) (20=5750) (21=6750) (22=8750) (23=14000) ///
  (24=27000) (50 90 99 = .), gen(v372m)
gen logv372m = ln(v372m)

* Staatsangehörigkeit
recode ef52 (1=1 "[1] Deutsch") (*=2 "[2] Ausländer"), gen(v52)

* sub: Bev. am Hauptwohnsitz, abh. beschäftigte Erwerbstätige ///
       (ohne Auszubildende) mit überw. Lebensunterhalt aus ///
       Erwerbstätigkeit und gültigen Bildungs- und Einkommensang.
gen sub = ef505>=1 & ef505<=2 & ef127>=4 & ef127<=6 & ///
          ef504==1 & ef338==1 & v372m<. & x287<. & x289<.

* (1) OLS (SRSWR)
xi: regress logv372m x287 x289 i.ef32 i.v52 if sub
estat imtest, white
```

```
estat hettest, iid

* (2) Robuste Varianzschätzung, Klumpung: Auswahlbezirk (SRSWR)
xi: regress logv372m x287 x289 i.ef32 i.v52 if sub, vce(cluster ef3)

* (3) SVY: Schichtung, Klumpung, Designgewichtung (SRSWOR)
gen schicht = ef1*10 + ef712
gen f = 0.01 * 0.035  // Auswahlsatz MZ 1%, CF 3,5%
gen w = 1/f           // Designgewicht
svyset ef3 [pw=w] , strata(schicht) fpc(f) singleunit(certainty)
xi: svy: regress logv372m x287 x289 i.ef32 i.v52 if sub

* (4) SVY: Schichtung, Klumpung, gebundene Hochrechnung (SRSWOR)
* Proxy Anpassungsschicht
gen anp=(1*(ef32==1 & ef52==1 & ef127~=9 & ef127~=10)) /// Deutsche Männer
      + (2*(ef32==2 & ef52==1)) /// Deutsche Frauen
      + (3*(ef32==1 & ef52~=1)) /// Ausländische Männer
      + (4*(ef32==2 & ef52~=1)) /// Ausländische Frauen
      + (5*(ef32==1 & ef52==1 & ef127==9)) /// Zeit-/Berufssoldaten
      + (6*(ef32==1 & ef52==1 & ef127==10)) // Wehrpflichtige
gen anpschicht=sub*(ef1*10+anp) /* nur für Subpop */
* Populationsdaten pro Anpassungsschicht (wie [A] in Beisp_06.do)
egen M_k = total(ef750g*100*(ef505<=2)), by(anpschicht)
svyset ef3 [pw=w] , strata(schicht) fpc(f) singleunit(certainty)
poststrata(anpschicht) postweight(M_k)
xi: svy: regress logv372m x287 x289 i.ef32 i.v52 if sub

exit
```